TIA 2011

9th International Conference on Terminology and Artificial Intelligence

Proceedings of the Workshops

WS1 — Terminological and ontological resources for extracting subjective information: how does ontology objectivity deal with sentiment subjectivity?

WS 2 — Ontology and lexicon: new insights

10 November 2011
INALCO
Paris, France

inalco
Institut national des langues et civilisations orientales

SORBONNE PARIS CITÉ
RECHERCHE ENSEIGNEMENT SUPÉRIEUR

IAOA

Workshop program

Thursday, November 10, 2011

TIA 2011 Workshops:

WS1 - Terminological and ontological resources for extracting subjective information: how does ontology objectivity deal with sentiment subjectivity?

WS 2 - Ontology and lexicon: new insights

9:00-9:15 Ouverture (Opening session)

Motivations du WS1 (WS1 motivations)

Terminological and ontological resources for extracting subjective information: how does ontology objectivity deal with sentiment subjectivity? — Introduction and Argument
Monique Slodzian

Motivations of WS2 (WS2 motivations): Goals of W3C "Ontology Lexica" community group — John McCrae

Ontology and Lexicon: new insights
Nathalie Aussenac-Gilles, Anne Condamines, Nathalie Hernandez and Bernard Rothenburger

9:15-11:00 Ressources termino-ontologiques pour l'extraction d'information subjective (Termino Ontological Ressources for subjective information extraction)

Les mots des sentiments : questions émergentes
Egle Eensoo-Ramdani

Construire un lexique translingue de sentiments à base de ressources existantes
Meng Sun

Subjectivité et sentiments : l'éclairage de la sémantique de corpus
Évelyne Bourion and Jugurtha Aït-Hamlat

Ontologies et folksonomies : même combat ?
François Rastier

Thursday, November 10, 2011 (continued)

Discussion (15 mn)

11:00-11:20 Pause café (Coffee break)

11:20-12:20 Applications (Applications)

Lexicalized ontology for the management of business rules. An industrial experiment
Nouha Omrane, Adeline Nazarenko, Peter Rosina, Sylvie Szulman and Christoph Westphal

Approach to the Creation of a Multilingual, Medical Interface Terminology
Joseph Roumier, Robert Vander Stichele and Laurent Romary

Ontology and Lexicon: The Missing Link
Fadi Badra, Sylvie Despres and Rim Djedidi

Discussion (Discussion) (15 mn)

12:20-13:50 Repas (Lunch)

13:50-14:50 Principes et théorie (Principles and theory)

Ontologies, Logic and Interaction: From Lexical Semantics to Geometrical Compatibility
Marco Romano

Towards an ontology based information system for Linguistic: the case study of the OTIM project
Julien Seinturier

On the ontological coherence of lexical resources
Laure Vieu

Thursday, November 10, 2011 (continued)

Discussion (Discussion) (15 mn)

14:50-15:50 Ressources Termino-Ontologique (Termino Ontological Ressources)

An Ontological and Terminological Meta-model for Semantic Information Retrieval
Axel Reymonet, Jérôme Thomas and Nathalie Aussenac-Gilles

Text-based IE and Open Linguistic Data for Termontological Resources
Andrés Domínguez Burgos, Koen Kerremans and Rita Temmerman

Ontology Lexicalisation: The lemon Perspective
Paul Buitelaar, Philip Cimiano, John McCrae, Elena Montiel-Ponsada and Thierry Declerck

Discussion (Discussion) 15 mn

15:50-16:10 Pause café (Coffee break)

16:10-17:25 Langue (Language)

Corpus-based extension of termino-ontology by linguistic analysis - a use-case in biomedical event extraction
Wiktoria Golick, Pierre Warnier and Claire Nedellec

Managing polysemy in the adventure tourism discourse with Frame Semantics
Isabel Durán-Muñoz

A natural language ontology-driven query interface
Enrico Franconi, Paolo Guagliardo, Sergio Tessaris and Marco Trevisan

Representing term variation in lemon
Elena Montiel-Ponsada, Guadalupe Aguado de Cea and John McCrae

Discussion (Discussion) (15 mn)

17:25-18:00 Discussion finale (Final discussion)

Table of Contents

Workshop Introductions

WS1 - Terminological and ontological resources for extracting subjective information: how does ontology objectivity deal with sentiment subjectivity?

WS 2 - Ontology and lexicon: new insights

Introduction and Argument

Monique Slodzian
ERTIM-INALCO
monique.slodzian@orange.fr

Subjective information extraction (opinion mining, sentiment analysis, subjectivity analysis) today is a key application of web-based information retrieval. This has given rise to such phrases as "sentiment word-list", "sentiment lexicon" or "sentiment ontology". In other words, epistemologically and linguistically speaking, one takes it for granted that there are specific word classes for conveying sentiments that are liable to be subsumed in an ontology. Yet, from the point of view of terminological doctrine, which was born from an alliance with science---the way it was received at the start of the 20th century---and lays claim to such features as rigour, consistency and systematicity, talking about sentiment terminology is a blatant oxymoron. The convergence between science and language was the result of a vital need---that of controlling scientific language. The gist of terminologies/ontologies was related to commanding scientific and technical vocabularies by accessing them in an onomasiological way--- from concept to word. The goal of making word formation more rational and operative went along with a relative depletion of syntactic expression, which was often turned into noun groups. The will to go beyond ordinary language, felt to be equivocal, unstabilised and fuzzy, demanded that the sentence be stripped of such items as adjectives, which are by nature subjective and emotional. The publication of the Vienna General Theory of Terminology in the early 1930s, setting out the programme of terminology, stems from the premise that scientific knowledge is based on logical reasoning and that its minimal unit – the "term" – is devoid of any overtone, has a single meaning, is accurate, refers to a single entity and can be included in an ontology. This is poles apart from ordinary language, which conveys the lowest level of abstract knowledge and is only capable of producing emotional and non-cognitive meanings. Opinion analysis, as it is done today, would simply not fit the bill.

Ever since it was set up 20 years ago, the TIA research group has challenged the positivist dogma. By using NLP tools, its members have contributed to shaking the belief that there is a gap between scientific and ordinary language. More generally, corpus analysis has weakened some certainties related to term definition, in particular the term-concept tie or the association of "concept" and "signifié".

The criticism of terminology that TIA-related researchers have embarked on was bound to take on the ontology movement whose excess is embodied in the limits of WordNet and, more recently, the ambitions of SentiWordNet.

Factoring in the historical framework briefly sketched out above as well as the epistemological directions and the practical achievements that have been part of TIA's activity from the start, we offer a workshop that will focus on such issues as:

- What is the standing of ontological-terminological sentiment resources? What is the nature of "concept-sentiments" (and their parts of speech)? What are the ties that connect the ontology's "concept-sentiments"? What are the relations (hypernym, meronym, etc) that can be built into a sentiment ontology? What would the root element be, which would subsume all others?

- Are ontological-terminological sentiment resources stable despite genre variation (review, forum post, blog post, blog comment ...)?

Workshop Introductions, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 1–2
Paris, 10 November 2011

- What is the part played by ontological-terminological sentiment resources in sentiment-detection applications? Do they play a key part? How do they combine with other linguistic levels (morphology, syntax)?

- In WordNet, synsets are the interlingua. What is the cross-language representation model offered in sentiment ontologies---such as SentiWordNet?

Ontology and Lexicon: new insights

Nathalie Aussenac-Gilles (IRIT, Toulouse) — Anne Condamines (CLLE-ERSS, Toulouse)
Nathalie Hernandez (IRIT, Toulouse) — Bernard Rothenburger (IRIT,Toulouse)

1 Ontologies and lexicons

During the last ten years, a large range of uses of ontologies, mainly user-interactive applications, access to document content, text analysis or information retrieval, have stressed the strength of the connection between ontologies and natural language. Firstly, textual resources can be used for learning ontologies from text, or at least to guide an interactive analysis of textual content to build ontologies. Secondly, ontology based semantic annotation and search, multilingual access to information, are some of the key features of the Semantic Web. One of the cornerstones then is to define the way in which knowledge models can be connected with their linguistic formulations. In particular, various studies defined representations of the lexical entries that contain information about how ontology elements (classes, properties, individuals etc.) are realized in multiple languages.

Considered as a knowledge representation, a lexicon gathers the linguistic properties of terms and their syntactic relations, whereas an ontology focus on a conceptual model. So the interface between these two layers is regularly discussed, in particular thanks to the notions of Lexical Ontology and Termino-Ontological Ressources (TOR). In both cases, the question is how to articulate knowledge models and their linguistic formulation, which refers to lexical entries and to their insertion in the discourse where they are used.

2 Relations within TIA, OntoLex workshops and the W3C "ontology and lexica" community group

In this scope, the OntoLex workshops have been organized. They made it possible to define research issues about the definition of rich and relevant models that can be used both to support linguistic processing based on lexical entities, and to provide an efficient access to knowledge representations. One of the key questions is about the joint representation of linguistic and conceptual information. These question turn out to be *"Which are the features required to represent lexical and ontological entities? How can these representation schemes be connected?"*. Possible answers are as diverse as the applications that include these models, the underlying semantic theories or the hypotheses related to the dynamics of meaning and interpretation.

Following these workshops, the "Ontology Lexica" W3C community group was born this year. Among its goals are not only the development of models for the representation of lexica (and machine readable dictionaries) relative to ontologies, but also the collection of best practices and experiment feed-back that demonstrate the added-value of such representations. The mission and research focus of the group are listed here:

`http://www.w3.org/community/ontolex/`

These research issues about ontologies and a lexicon are very relevant for the TIA conference, that, up to now, has been promoting innovating works in terminology, questioning the positivist view that

Workshop Introductions, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 3–4
Paris, 10 November 2011

freezes the representation of meaning and its interpretation. The TIA conference questioned in particular the notion of terminological and ontological resource (TOR), which is close to representations like ontologies with a lexical component.

Ever since it was set up 20 years ago, the TIA research group has challenged the positivist dogma. By using NLP tools, its members have contributed to shaking the belief that there is a gap between scientific and ordinary language. More generally, corpus analysis has weakened some certainties related to term definition, in particular the term-concept tie or the association of "concept" and "signifié".

3 Topics

The talks given during this workshop address issues in the following, non-exhaustive list:

- presentation of models and representations that combine ontologies with lexical or terminological entities;
- use-cases that illustrate the added-value of these representations inside software applications;
- notion of "knowledge rich context" and how it can be taken into account in knowledge representation;
- ability of these representations to improve linguistic analyses and language processing or access to ontology entities;
- ability to use these representations to account for polysemy, meaning change over time, or knowledge evolution.

Les mots des sentiments : questions émergentes

Egle Eensoo-Ramdani
ERTIM-INALCO

Depuis une dizaine d'années, l'analyse automatique de l'expression de l'évaluation dans le discours, ou sentiment analysis, s'est développée fortement en raison de ses nombreuses applications potentielles. La constitution d'un lexique approprié — les « mots des sentiments » — a été une des premières approches pour accéder aux sentiments. Même si d'autres approches, plutôt quantitatives ont rapidement émergé, l'idée d'une ressource de qualité qui permettrait d'analyser finement des textes est toujours présente. Cela soulève de nombreuses interrogations. Les exigences des ressources terminologiques et ontologiques (stabilité, objectivité, référentialité) sont-elles compatibles avec la nature subjective de ces éléments ? Que veulent dire les caractéristiques très utilisées comme la positivité et négativité ? L'expression des sentiments dans un texte est-il une affaire du lexique ?

Construire un lexique translingue de sentiments à base de ressources existantes

Meng Sun
SEDYL-INALCO

In the domain of sentiment analysis, numerous works on the construction of lexicon of emotion have been performed for English language. In this situation, we would like to create a similar lexicon for the French language. In stead of building a monolingual corpora and then extract a word list from it, we have experimented with a translingual method. We have firstly gathered certain reliable and large-scale linguistic resources, such as WordNet, SentiWordNet, French WordNet. Each of them has some specific features favorable for our project. Then, we've borrowed necessary information from each of them and build our own lexicon, a French SentiWordNet, in several steps. In this report, we are going to present the features of those resources and the methodology of our work.

Subjectivité et sentiments : l'éclairage de la sémantique de corpus

Évelyne Bourion **Jugurtha Aït-Hamlat**
ERTIM-INALCO

L'analyse des contextes exprimant la peur dans deux corpus différents par le discours, le genre textuel et la période (roman du discours littéraire des XIX et XXè siècles vs billet du témoignage autobiographique en ligne du XXIè siècle) permet de mettre en question les présupposés que tout locuteur maîtrisant une langue serait amené à formuler. En effet, « mettre en texte » ce n'est pas seulement employer les « mots de la peur » (des formes verbales comme j'ai peur ou des substantifs comme terreur, effroi, crainte, etc.), c'est donner à reconnaître des « formes sémantiques » partagées dans une aire culturelle. Ainsi l'expression textuelle des sentiments et de leur intensité rejoint, au-delà des variations de genre, discours, époque, les savoirs anthropologique et psychologique plutôt que le discours de la logique.

Ontologies et folksonomies : même combat ?

François Rastier
ERTIM-INALCO

Les ontologies sont réputées représenter des faits, les folksonomies des valeurs ; mais les évaluations sont diffusées partout, dans tout langage même contraint, et même en-deçà du palier lexical. On distinguera a minima les émotions (de base), les évaluations (traits sémantiques), et les sentiments (formes sémantiques complexes) ; enfin les régimes textuels de subjectivisation ou d'objectivation. Les métadonnées évaluatives doivent être problématisées selon les champs d'application : page-ranking, recherche d'information, branding, création d'en-têtes headers ou de profils, selon qu'on indexe les documents par les individus ou l'inverse. Sans angélisme technologique, on peut distinguer des enjeux de pouvoir complémentaires : militaire, politique et administratif pour les ontologies ; économique pour les folksonomies.

Lexicalized ontology for the management of business rules
An industrial experiment

Nouha Omrane* **Adeline Nazarenko*** **Peter Rosina**** **Sylvie Szulman*** **Christoph Westphal****
*LIPN UMR 7030 (Université Paris 13 & CNRS), France
firstname.lastname@lipn.univ-paris13.fr
** AUDI AG, Germany
firstname.lastname@audi.de

Abstract

This paper[1] describes a simple formalism designed to encode lexicalized ontologies and shows how it is used in a business rule management platform[2] of the automotive domain.

1 Introduction

Business Rules Management Systems (BRMSs) are software applications that help organizations to separate their application code from their business knowledge. BRMSs help the users to author and maintain business rules and apply decision logic that reflects this business knowledge. However, domain experts who are not always business rules experts may have difficulties expressing their knowledge in formalized logic languages. Supporting them in their management of the knowledge needed to write these rules is one of the goals of the ONTORULE project.

We propose building an ontology as a formal model for representing conceptual vocabulary that is used to express business rules in written policies. OWL-DL language is used to represent concepts and properties of the domain ontology but such an ontology must be linked to the lexicon used to express rules in the text, so experts can query source documents. This calls for a formalism to link linguistic elements to conceptual ones. We opt to use the SKOS[3] language which provides basic elements to link domain concepts to terms from the text. The combination of OWL entities, SKOS concepts and their related information form a lexicalized ontology which supports the semantic annotation of documents.

The paper is organized as follows. Section 2 describes the Audi use case, on which this approach has been tested. Section 3 explains the choice of SKOS combined with OWL as language to support the lexicalized ontology. Section 4 reports the experimentations made in the Audi use case.

2 The Audi use case

Nowadays, the development of new cars has become very challenging and many different process steps are involved. Computer Aided technologies, like virtual modeling, simulations or the analysis and planning of physical testing, need to be integrated even tighter to satisfy the higher requirements and reduced time-to-market which also shortens the development cycles.

In the ONTORULE project, Audi is developing a prototype BRMS that makes use of ontologies and business rules. Ontologies together with business rules help Audi to keep abreast of technology advances and use them in its R&D IT applications. Especially the interweaving of the various Computer Aided technologies will help Audi to reduce development time and cost.

One of the difficulties with business knowledge rules is that various departments or roles sometimes use different vocabularies for the same things so they cannot understand each other immediately. Addionally, formalized rules *per se* are often not easy to understand. Using an ontology as a unified model for a heterogeneous vocabulary will reduce misunderstandings and ensure that people are discussing the same thing. Also, the users can easily confirm and verify the appropriateness of

[1] This paper is an extract from our paper (Omrane et al., 2011a)

[2] This work was realised as part of the FP7 231875 ONTORULE project (http://ontorule-project.eu). We thank our partners for the fruitful discussions, especially to Audi for the collaboration on their use case.

[3] Simple Knowledge Organization System

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 9–12
Paris, 10 November 2011

the modeled semantic relations. Finally, the prototype that is to be developed is expected to handle links between source documents, such as policies or internal documents, and the concepts and instances of the ontology.

3 A formalism for lexicalized ontologies

3.1 Existing formalisms

Many research activities have tackled the problem of linking an ontology to a lexicon. Two major areas are of interest. The first is the NLP domain which aims at adding some semantic structure to a lexicon by linking its elements to ontology's elements. There are several ways to combine a lexicon with an ontology: *LMF*[4] standard (Francopoulo et al., 2007), *TMF*[5], *OLIF*[6], *LMM*[7]. The other group trys to link an ontology to a lexicon by modeling linguistic information in the ontology as in (Reymonet et al., 2007), *LexOnto* (Cimiano et al., 2007) or *LIR* (Peters et al., 2009). There also exist more abstract approaches like *LingInfo* (Buitelaar et al., 2006), which defines a meta class to link the linguistic properties to the concept or to its Data/Object properties, or (Ma et al., 2009), which introduces a set of annotation rules to link an existing ontology to its lexicon.

From a practical point of view, the choice of one model or another depends on the aimed application and the task. Our aim is to build a lexicalized ontology to allow annotating the technical documents and thus to help the expert in exploring documents by querying its set of annotations. We use for that *SKOS* W3C standard that links linguistic to semantic knowledge.

3.2 A SKOS-based approach

A key issue for experts in managing a rule base is to be able to mine textual sources to understand how a given concept is used in business documents, what rules are related to it and how those concepts and rules evolve when the policies are updated. This is achieved through the semantic annotation of the documents in which the mentions of the ontological entities (concepts, instances and roles) are highlighted and can be searched for.

[4]Lexical Markup Framework

[5]Terminological Markup Framework

[6]Open Lexicon Interchange Format

[7]Linguistic Meta Model

Our aim is therefore to save the terms related to the conceptual vocabulary that is used to express the business rules. We don't need to encode sophisticated information such as the morphological structure of terms since we do not perform a deep analysis of the documents. We simply need to save the various linguistic units that denote a concept, instance or role. SKOS supports encoding of *SKOS concepts* that represent the links between the OWL concepts and their related terms, which are encoded as *skos labels*[8]. This relation is described by `<rdf:Description rdf:about>`.

When designing and updating business rules, experts face the problem of the heterogeneity of information sources and multilingualism. *SKOS* also supports that normalization of vocabularies. A given SKOS concept can be associated with the various terms or labels that denote it in the texts or any other information source. For a given concept, *SKOS* supports distinguishing one preferred label and as many alternative labels as necessary, using the `<skos:prefLabel>` and `<skos:altLabel>` properties. In the Audi ontology, for example, the SKOS concept *LowTemperatureChamber* is linked to two terms: *low temperature chamber* is encoded as the preferred label and *refrigerated cabinet* as its alternative form.

SKOS also supports the encoding of multilingual information. The information about the language used is described by `<rdf:lang=''en''>`. For example, the SKOS concept *TrolleyTest* has a preferred label "trolley test" which is mentioned in English texts, and an alternative label "Schlittentest" in German.

Since experts often have to manage a large volume of information but do not always formally describe all the concepts, it is important to add informal documentation when it is available. Defining concepts in natural language is very important to understand what concepts mean, especially if they have ambiguous or implicit labels. Those definitions can sometimes be extracted from the source documents when designing the ontology. In that case, they are associated to the related SKOS concepts using the label `<skos:definition>`.

In such a lexicalized ontology, the domain concepts and their occurrences in the text can be

[8]http://www.w3.org/2004/02/skos/

matched from one to another thanks to the linkage of OWL entities, SKOS concepts and labels. This is a simple efficient way to represent lexicalized ontologies and we show in the following section its benefit for the Audi BRMS. Figure 1 describes how the Audi ontology is linked to the lexicon and annotated text.

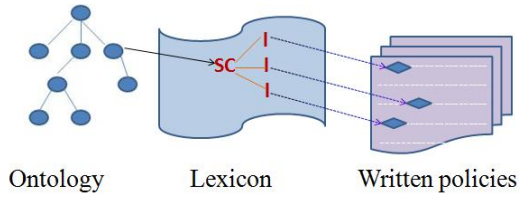

Figure 1: A lexicalized ontology for annotating source documents. Each concept from the ontology is linked to a SKOS concept *SC* and each SKOS concept is related to its labels. The annotations link some text entities to these labels

4 The Audi lexicalized ontology

This section presents the Audi ontology and illustrates the benefit in the Audi use case of having such a lexicalized ontology.

That ontology has been built in two steps. At first, the goal was to integrate the various existing knowledge sources in a single one. This resulted into a small conceptual model (around 30 concepts) associated with a large knowledge base (thousands of instances).

In a second step, in order to better fits the experts' needs for semantic querying and document mining, the initial ontology has been restructured and lexicalized. It also appeared useful to increase the granularity of the domain model so as to represent for instance not only the various types of tests but also their actual occurrences in the car manufacturing process (instances that are related to the different tests applied to specific vehicle models).

This led to encoding various elements as concepts rather than instances (90 concepts were added). The conceptual structure has been reorganized (4 subsumption levels instead of 1). A SKOS resource has been associated with this resulting ontology: each concept is related to at least 1 preferred label and up to 5 alternative labels. In addition, using a subset of the initial ontology for the exploration of written policies showed that some of the mentioned concepts were missing in the initial ontology and led us to enrich it (Omrane et al., 2011b).

Once the ontology is lexicalized, domain experts can query source documents to search for fragments of texts that describe specific concepts mentioned in rules. For example, they can find all references of the concept *BreakingStrengthOfStrapTest* in the text, wherever it is mentioned in the documents. They can also search for all sentences where the physical methods are mentioned in the text. As the concepts expressing tests are subconcepts of the concept "MethodInformation", we query the text by searching about all labels describing subconcepts of "MethodInformation".

Thanks to the labels of concepts, the ontology can be used to annotate the documents. Figure 2 shows an example of texts where all the mentions of known concepts are emphasized. This supports experts in browsing of documents.

Two belts or restraint systems are required for the buckle inspection, the low-temperature buckle test, the low-temperature test described in paragraph 7.5.4. below where necessary, the buckle durability test, the belt corrosion test, the retractor operating tests, the dynamic test and the buckle-opening test after the dynamic test. One of these two samples shall be used for the inspection of the belt or restraint system.

Figure 2: A fragment of text annotated by the lexicalized ontology.

5 Conclusion

The proposed integration of Computer Aided technologies will increase the flexibility of the development process, allowing Audi to meet the increasing market demand for product diversification. This integration relies on the design of an application that is currently under development and is based on a BRMS.

Our approach for the acquisition and management of the knowledge embodied in such BRMS relies on a lexicalized ontology which unifies and normalizes the various vocabularies and links the conceptual knowledge to the source policies and regulation written in natural language. Using a lexicalized ontology enables experts to determine the most suitable Computer Aided technologies from given functional requirements and to query sources documents.

These new approaches, standards and technologies are already partially integrated in some processes. During the next years Audi will continue to incorporate the ONTORULE platform in their landscape which will lead to even less time-

consuming, cheaper and higher quality processes in the innovation and development cycles.

References

Paul Buitelaar, Michael Sintek, and Malte Kiesel. 2006. A Multilingual/Multimedia Lexicon Model for Ontologies. In *ESWC*, pages 502–513.

Philipp Cimiano, Peter Haase, Matthias Herold, Matthias Mantel, and Paul Buitelaar. 2007. LexOnto: A Model for Ontology Lexicons for Ontology-based NLP. In *Proceedings of OntoLex - From Text to Knowledge: The Lexicon/Ontology Interface (workshop at the International Semantic Web Conference)*.

Gil Francopoulo, Nuria Be, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2007. Lexical markup framework: ISO standard for semantic information in NLP lexicons. In *Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV*.

Yue Ma, Laurent Audibert, and Adeline Nazarenko. 2009. Ontologies étendues pour l'annotation sémantique. In *20mes Journées Francophones d'Ingénierie des Connaissances (IC09)*.

Nouha Omrane, Adeline Nazarenko, Peter Rosina, Sylvie Szulman, and Christoph Westphal. 2011a. Lexicalized ontology for a business rules management platform: An automotive use case. In *RuleML@BRF*, Florida, États-Unis, November.

Nouha Omrane, Adeline Nazarenko, and Sylvie Szulman. 2011b. Les entités nommées : éléments pour la conceptualisation. In *22mes Journées Francophones d'Ingénierie des Connaissances (IC11)*.

W. Peters, M. Espinoza, E. Montiel-Ponsoda, and M. Sini. 2009. Multilingual and localization support for ontologies. Technical report, D2.4.3 Neon Project Deliverable. Technical report.

Axel Reymonet, Jérme Thomas, and Nathalie Aussenac-Gilles. 2007. Modélisation de ressources termino-ontologiques en OWL. In Francky Trichet, editor, *Journées Francophones d'Ingénierie des Connaissances (IC), Grenoble*, pages 169–180, July.

Approach to the Creation of a Multilingual, Medical Interface Terminology

Joseph Roumier
Heymans Institute of Pharmacology
/ Ghent, Belgium
CETIC / Charleroi, Belgium
Joseph.Roumier@cetic.be

Robert Vander Stichele
Heymans Institute of Pharmacology
/ Ghent, Belgium
Robert.VanderStichele@ugent.be

Laurent Romary
INRIA & HUB-IDSL
/ Paris, France
Laurent.Romary@inria.fr

Abstract

When confronted with registering and searching for medical information, health workers perform poorly. The various outputs of such situation are frustration all along the information line, missed opportunities of accurate treatment, suboptimal health policies, failed clinical trials. As a result an easier management and a simplified consultation of the information resources for health professionals and patients is a key issue.

The naive approach that would consist in standardizing the languages and terms used is not acceptable. These different terms have many reasons to be valid and to seek unity of language and words would also be detrimental to the health of patients.

Instead of trying to over-simplify the problems we should accept them even if it means more complicated systems. We must take into account that there are vocabulary differences between Specialists and General Practitioners talking about the same medical fact. There are even more differences between the patients and the doctors. Also, the vocabulary being used evolves over time and space and many local expressions exist to designate the same diseases, body parts.

In order to cope with this heterogeneity in a(n) (semi)automated manner, in fact to perform the task of the doctor being the interface between the medical wor(l)d and the lay person wor(l)ds, natural language processing is necessary. Even though some mechanisms exist, the effort to maintain a central and evolving multilingual terminology containing all the linguistic intricacies and the local lexical variants for a concept would be daunting.

That is why we propose a terminological system that contains two types of domain-specific resources.

- First, a reference interface terminology[Rosenbloom et al., 2006], [Rosenbloom et al., 2008], multidisciplinary, multilingual but containing only the reference concepts and their standardized as well as local lexical representation. These reference concept are linked to nomenclature, such as SNOMED-CT[1], or bibliographic thesauri such as Medical Subject Headings, or to international classifications .
- The second resource, is a series of specific, lexical and often uni-lingual end-user terminologies that must be linked to the multilingual reference terminology. These lexicons can be linked to Natural Language Processing applications, and either be oriented to patients or either to professionals (e.g. local nomenclatures).

We propose a dual mechanism to link the first type of resource – the reference terminology - and the second type of resources – the end-user monolingual terminologies:

- The concept in the reference terminology is linked to the sense part of a lexical

[1] http://www.ihtsdo.org/snomed-ct/

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 13–15
Paris, 10 November 2011

resource. This mechanism preserves the conceptual integrity.

- The alignment of the lexical representation of the concept in a specific language with the corresponding lemma in a lexicon of that language.

Tthe first type of resource is created using the association of the Terminological Markup Framework (TMF) [ISO 16642, 2003], [Romary, 2010]) as the meta-model and a carefully chosen subset of the data-categories found on the ISOcat.org platform [ISO 12620, 1999]. The resulting Terminological Markup Language (TML) (see Illustration 1) is serialized using RelaxNG[2]. For this part of the work, we were inspired by the TML created by the TermSciences[3] project [Khayari et al., 2006].

For linking this new resource for semantic interoperability; we chose for as the backbone of the interface terminology SNOMED-CT[4], but also MeSH[Lipcomb, 2000], and a series of other external classifications to link to, such as, ICD-10[ICD-10, 2010], ICPC [ICPC-2-R, 2005], the International Classification for Primary Care, LOCAS[Jamoulle & Roland, 1993] which is a General Practitioner oriented resource.
For the second resource our proposal is to use the Lexical Markup Framework (LMF)[ISO 24613, 2008], [Romary, 2010] for that purpose because it is conceived to deal with linguistic intricacies, and uses the same set of linguistic ISOCat.org[5] Data Categories. LMF provides guidelines on how to link its entries with TMF and other concept based representation systems. Finally the meta-model contains a mechanism to deal with multiple senses. We were inspired for the relationship between lexicons and ontologies by [Cardillo, 2011].

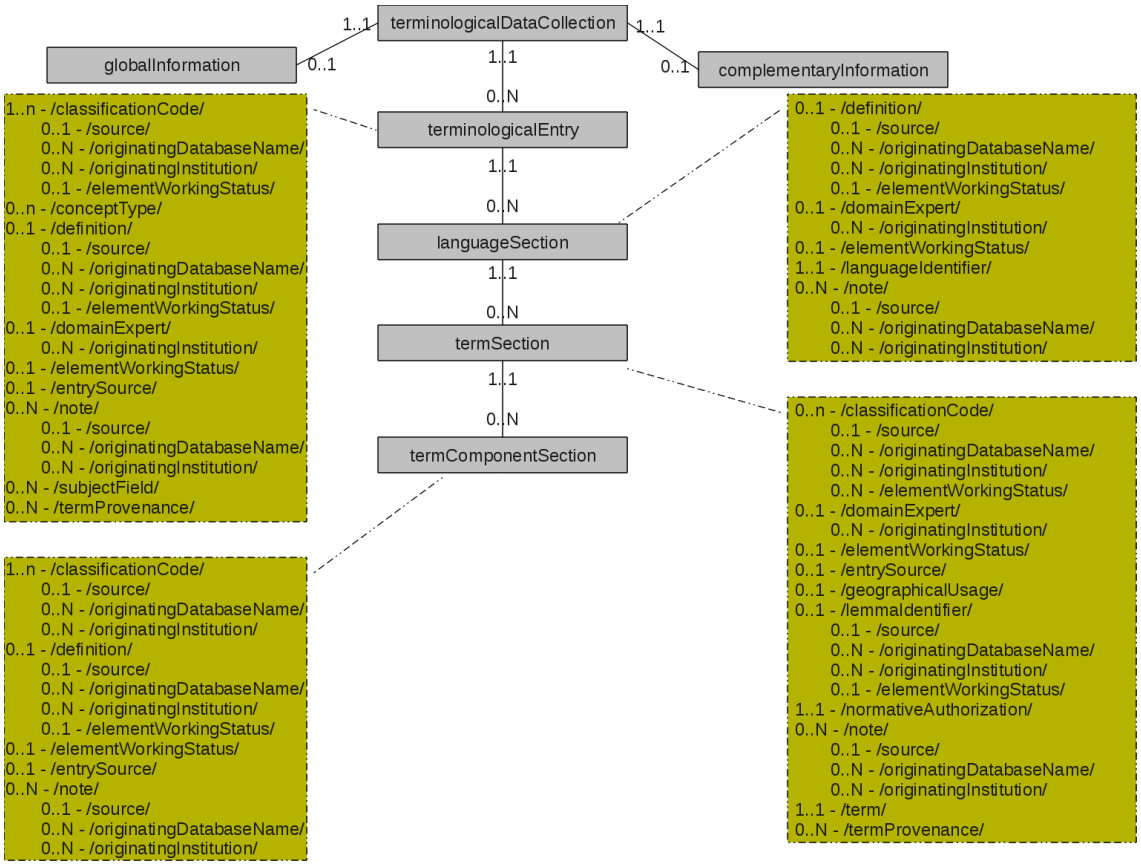

Illustration 1: TMF and Data Categories for the main levels

[2] http://www.relaxng.org/
[3] http://www.termsciences.fr/
[4] http://www.ihtsdo.org/snomed-ct/
[5] http://www.isocat.org/

In addition to this it is worth mentioning that the work is being done with the idea of publishing the resources online following the Linked Data[Bizer et al., 2010] principles produced by the Semantic Web initiative of the World Wide Web Consortium(W3C)[6].

In conclusion the work relies on three pillars: first the existence of an ISO standard to develop models for multilingual terminologies that take into account the diversity of terminology sources while preserving interoperability and sustainability. Second, the existence of an ISO standard to develop models for mono- (and multi-) lingual lexicons, that tap into the existing body of language-specific linguistic resources. Third, the existence of a W3C standards for the publication of semantic data on the Internet. The work is ongoing and currently being reviewed by the health and environment department of the Belgian government.

References

[Bizer et al., 2010] Bizer, C. and Heath, T. and Berners-Lee, T., Linked data-the story so far, sbc, S.9, 2010

[Cardillo, 2011] Cardillo E. A lexi-ontological resource for consumer healthcare. The Italian Consumer Medical Vocabulary. [Doctoral Thesis]. Fondatione Bruno Kessler. April 2011.

[ICPC-2-R, 2005] ICPC-2-R - Wonca (World Organisation of Family Doctors) ICPC-2-R, International Classification of Primary Care (revised 2nd Ed). OUP. 2005

[ICD-10, 2010] ICD-10 -International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Version for 2007.Tabular List of inclusions and four-character subcategories- http://apps.who.int/classifications/apps/icd/icd10online/

[ISO 16642, 2003] ISO 16642, Computer applications in terminology - Terminological markup framework (TMF), 2003.

[ISO 12620, 1999] ISO 12620:1999, Computer applications in terminology – Data categories, 1999.

[ISO 24613, 2008] ISO 24613:2008, Language resource management — Lexical markup framework (LMF)

[Jamoulle & Roland, 1993] « LOCAS-CISP, logiciel de codage et d'acquisition de synonymes pour la Classification Internationale des Soins Primaires » (Jamoulle M, Roland M Fédération des Maisons Médicales, Bruxelles), 1993.

[Khayari et al., 2006] Khayari, Majid and Schneider, Stéphane and Kramer, Isabelle and Romary, Laurent, Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative., 2006, http://hal.archives-ouvertes.fr/hal-00022424

[Lipcomb, 2000] Lipscomb, C. E., Medical subject headings (MeSH), Bulletin of the Medical Library Association, S.265, 2000

[Romary, 2006] Romary, Laurent and Kramer, Isabelle and Alt, Susanne and Roumier, Joseph, Gestion de données terminologiques : principes, modèles, méthodes, S.13, 2006, http://hal.archives-ouvertes.fr/hal-00096910

[Romary, 2010] Romary, Laurent, Standardization of the formal representation of lexical information for NLP, 2010, http://hal.inria.fr/hal-00436328

[Rosenbloom et al., 2006] Rosenbloom ST; Miller RA, Johnson KB, Elkin PI, Brown SH. Interface terminologies: Facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13:277-288.

[Rosenbloom et al., 2008] Rosenbloom ST, Miller RA, Johnson KB, Elkin PI, Brown SH. A model for evaluating interface terminologies. J am Med Inform Assoc 2008;15:65-76.

[6] http://w3.org/

Ontology and Lexicon: The Missing Link

Fadi Badra **Sylvie Despres** **Rim Djedidi**
LIM&BIO
Université Paris 13
UFR de Santé, Médecine et Biologie Humaine (SMBH) - Léonard de Vinci
74, rue Marcel Cachin 93017, Bobigny Cedex France
fadi@fadi.lautre.net sylvie.despres@univ-paris13.fr rim.jedidi@univ-paris13.fr

1 Introduction

Ontologies specify formally concepts and relations of a specific domain and their related constraints (axioms, rules, etc.). Lexica (or terminologies) define terms that refer to a domain as lexical entities associated to linguistic information (morpho-syntactic properties and linguistic relations between terms).

Ontology building from text methodologies deal with intermediate representation levels associated to the different available resources. These intermediate levels –including lexica and termino-ontological resources – contain rich linguistic information on the initial corpus that is lost in the formal representation of the final ontology. In fact, transition from lexical layer to ontological layer looks like a sleight of hand. It is driven implicitly as it is buried in ontologist's mind and no trace of the activity remains in the resulting ontology. This is the *missing link* between lexical and ontological layers. It is not possible to represent everything either in lexicon or in ontology. We need an interface that keeps the link between the two layers. And for that, we need to think about the format of such interface.

Several researches (Szulman et *al*., 2009; Tiscornia, 2006; Cimiano et *al*., 2011) have underlined the interest of preserving the link between lexical and ontological layers and articulating linguistic expression with the associated knowledge model. Moreover, emerging initiatives[1] are working on defining a representation model that ensures the interface between these two layers.

The paper is structured as follow: first, we summarize an experience feedback in nutrition domain through two use cases and then, we discuss raised points and conclude.

2 Experience Feedback in Nutrition Domain

Working on knowledge modelling and exploitation in nutrition domain and on recipe corpus analysis, we have been confronted to the need of articulating ontologies with their associated lexical components and thus, to the need of a representation model carrying out this articulation.

In this section, we summarize through two use cases, the raised issues and the alternatives adopted in this work.

2.1 Use case 1: a lexicon and an ontology for recipe search engine

In this use case, we have worked on improving a recipe search engine (Benamar, 2011). A lexicon has been built from recipe corpus to facilitate user query analysis. Then, an ontology has been developed to complete the bringing-in of the lexicon by providing reasoning capabilities to the search engine.

The approach adopted combines syntactic and semantic methods. A first set of domain nutrition knowledge and descriptive properties has been identified and exploited to help in selecting relevant recipes.

Recipe corpus analysis with NLP tools (Ogmios platform, TreeTagger, Yatea and SynoTerm) has produced a list of linguistic entities – associated to their lemma, grammatical category and synonyms – structured in an XML

[1] http://www.w3.org/community/ontolex/

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 16–18
Paris, 10 November 2011

lexicon. The lexicon is composed of terms related to ingredients, quantities, unities, kitchenware, preparation methods, etc. It brings a first level of enhancement to the search engine taking particularly into account term synonymy, and hyperonymy to distinguish terms designating specific notions from those designating general ones (as "monkfish" and "fish").

The aim of this work was also to provide results that are suitable to user research criteria, profile and preferences, and that give nutritional recommendations. As the lexicon is not enough to meet this purpose, we have developed an ontology that models the vocabulary associated to recipes and integrates nutritional and physiopathological knowledge.

This use case has confirmed the complementarity between lexicon and ontology and the importance of the link between them. Lexicon provides a linguistic base that allows user query analysis and facilitates extraction of recipe results. Ontology provides a semantic referent that enables reasoning mechanisms to enrich user query and extract the most suitable recipes.

To extend engine usage by including in the research recipes available on the web, it is necessary to translate imported recipes in a format that is compatible with the built ontology. Processing these recipes also needs to exploit the lexicon. XML format representing the built lexicon do not allow to fully exploit linguistic information that can be associated to ontology. A more expressive format would be more appropriate for this need.

The second use case presented in the following section, underlines the interest of using a rich lexicon representation format.

2.2 Use case 2: information extraction from recipe corpus guided by a lexicon and an ontology

A model as LEMON (McCrae et *al*., 2011(a)) provides a rich expressivity to lexicon representation associating to each lexical entry a lemma, a lexical form, components, and also a lexical sens ensuring the link with the associated ontological reference (Mccrae et *al*., 2011(b)).
Some ongoing work aims at comparing LEMON with conventional ontology lexicalization approaches in the context of ontology-based information extraction (IE).

In (Davis et *al*., 2011), LEMON was exploited to automatically generate lexical resources associated with a cooking ontology and the resulting ontology was used to semantically annotate a small text corpus of 4650 lines of cooking recipes. The study showed that the LEMON API can be easily wrapped as a resource in the open-source text analysis framework GATE (Cunningham et *al*., 2011) to write ontology-based gazetteers that exploit LEMON-generated lexical resources.

A first version of a LEMON gazetteer (called the LemonOntoGazetteer) has been implemented in GATE and its performance was compared with the state of the art (OntoRooGazeteer[2]) that is already available in GATE. While this work constitutes only a preliminary study, the first experiments were encouraging since the LemonOntoGazetteer matched 74% of the 798 annotations created by the OntoRootGazetteer.

More research is however needed in order to study more thoroughly the benefits of using LEMON for ontology lexicalization in the context of ontology-based IE. In particular, the lexical model generated by LEMON served only in a pre-processing phase to generate a list of entries to be used by a conventional list gazetteer.

Future work will include writing a full-blown LEMON Gazetteer for GATE that exploits LEMON as a lexicon model during the entity recognition phase. Besides, the performance of the LemonOntoGazetteer was only compared to the output of the OntoRootGazetteer. Running more thorough experiments will require to create a Gold Standard of cooking recipes semantically annotated by domain experts on which the performance of the LemonOntoGazetteer could be compared with other semantic annotation processes.

3 Discussion and Conclusion

Through these use cases, it appears necessary to have both a formal ontology as a semantic referent, and a lexicon as a rich linguistic base represented in an expressive format. We also need a representation format of linguistic information that preserves richness of the exploited terminological resources.

We might think to represent everything in the ontology (lexical and conceptual aspects) but this inevitably deprives lexical aspect and limits the evolution of domain lexicon. We might also

[2] http://gate.ac.uk/sale/tao/splitch13.html#sec:gazetteers:ontoRootGaz

think to transform existing lexical resources by using standards as RDFS and SKOS, but this solution even if it could be applied to some resources, limits linguistic information associated to ontology (Tian, 2011; Mccrae et *al.*, 2011(b)). For some other resources, it is not even clear how to translate them automatically in these standards.

Complementarity between lexicon and ontology is thus obvious. To take advantage from it, we need to define a representation format that allows articulating lexical and ontological layers. It would be a first step to fill the *missing link*.

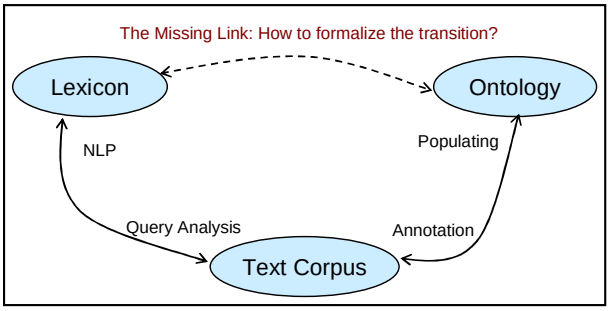

Figure 1. Ontology and Lexicon: The Missing Link

References

Benamar S. 2011. Toward integrating a knowledge layer to a research engine in nutrition domain (in French). Bioinformatics Master Dissertation. Paris 13 University.

Brian D., Badra F., Buitlelaar P., Handschuh S., Wunner T. 2011. Squeezing LEMON with GATE. The 10th International Semantic Web Conference ISWC 2011, Bonn Germany.

Cimiano P., Buitelaar P., McCrae J., Sintek M. 2011. LexInfo: A declarative model for the lexicon-ontology interface J. Web Sem. 9(1): 29-51

Cunningham, H., et *al*. 2011. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. ISBN 0956599311.

McCrae J., Aguado-de-Cea G., Buitelaar P.,Cimiano Ph., Declerck T., Gómez Pérez A. Gracia J., Hollink L., Montiel-Ponsoda E., Spohr D., Wunner T. 2011 (a). The Lemon cookbook. Monnet project.

McCrae J., Spohr D., Cimiano P. 2011(b). Linking Lexical Resources and Ontologies on the Semantic Web with lemon. Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Crete.

Szulman S., Charlet J., Aussenac-Gilles N., Nazarenko A., Sardet E., Teguiak H.V. 2009. DAFOE: an Ontology Building Platform From Text or Thesauri. In International Conference on Knowledge Engineering and Ontology Development (KEOD 2009).

Tian T. 2011. Identification And Analysis Of Lexical Resources In Nutrition Domain To Be Translated In SKOS (in French).

Tiscornia D. 2006. The LOIS Project: Lexical Ontologies For Legal Information Sharing. Proceedings of the V Legislative XML Workshop.

Ontologies, Logic and Interaction: From Lexical Semantics to Geometrical Compatibility.

Marco Romano
Department of Philosophy – Università Roma Tre, Roma, Italy
mromano@uniroma3.it
and LIPN – Université Paris 13, Villetaneuse, France
marco.romano@lipn.univ-paris13.fr

Abstract

We introduce a logical model to represent ontologies as well as folksonomies in the Web by means of geometrical objects called Compatibility Spaces. Besides its generality – which makes it suitable both for ontologies and folksonomies – this approach provides also some intuitions concerning the question on the nature of the fundamental elements both in semantic lexica and linguistic ontologies, the main issues sounding like 'are these elements either proper to the nature of the world or of the language?' and 'can we have any ontology without language?'.

1 Organizing knowledge in the Web

There exist different approaches to ontologies, even leaving aside all the history of philosophical ontology and focusing just on ontologies for use in some information system (in a very broad sense). We consider in particular the use of ontologies to the aim of building the Web of data, that is as a key component towards Semantic Web. In this context, ontologies are to describe, and possibly define, what the resources in some Web repository are (be they low level data, multimedia resources, text documents ...) and can be used for. According to the particular environment where the ontology is to be used, it can be designed in different ways. For sake of simplicity – and also in order to roughly identify two somehow contrasted positions – we mention just a couple of them. One is rooted in the long-lived practice of semi-automatic linguistic analysis by means of NLP techniques and leads to thesauri, vocabularies, or even lexical ontologies whose backbone is a taxonomy of words / concepts resulted as most relevant (by the number and place of occurrences) in the examinated *corpora* – after an important work by human experts in assessing the result of machine-operated analysis. Typically, ontologies designed this way are used to classify resources, which means in particular to say what a text document is about (be it a blog entry, forum post, email message or a paper, internal documentation ...) within a well defined context (an organization, a web portal).

The other approach is (or should be) more committed to the special context of the World Wide Web and exploits at its most the 'linked data' paradigm. A thorough survey, analysis and formal modeling of the domain of application of the to-be-designed ontology is conducted by Knowledge Representation experts assisted by domain experts. The domain is somehow re-engineered from scratch by creating an ontology that either is also richly connected with other widely accepted ontologies and/or vocabularies – hence also contributing to the deployment of the 'Web of ontologies' – or is rather isolated but very finely formalized, allowing even for automatic reasoning. Such a task-specific design of ontologies leads to ontologies able to describe functions, processes and activities of the adopting organization, besides what the single resources are.

To be honest, we must admit that actually these two approaches are not two worlds apart. Concept hierarchies for engineers' taxonomies often come from semi-automatic *corpora* analysis. And the possibility itself to do semi-automatic ontology extraction from text analysis over a *corpus* relies on the previous definition of a small set of basic categories and concepts that bring necessarily some foundamental ontological assumptions in the process of deriving the ontology from the analysis of the language used to talk about a given domain.

Finally, we must also consider another way of organizing information over the Web: tagging. It is the well-known attitude of users of Web2.0 services who not only provide on their own data and contents for the Web, but also produce some kind of organization of the Web content by tagging resources, thus enabling a dirty-but-working form of keywords search. Collections of tags can hardly be compared to 'true' ontologies. Nevertheless if one considers the set of tags collections used by all

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 19–21
Paris, 10 November 2011

users in a given community – i.e. a folksonomy – s/he will be able to discover some organizations of resources that may provide interesting insights concerning relationships between the concepts that are relevant to a given resource, and the words and terms used to record them in the tag.

2 Substantial differences

Whereas the ideal way to Semantic Web would require to eploit the best from each approach by mixing them all together, we observe that a number of substantial differences characterize these approaches, so that their actual combination looks really problematic.

The linguistic approach, in principle, might deal with the whole World Wide Web in general: Imagine of a very large lexical ontology, possibly resulting from linking to each other many smaller ontologies. Thus, this approach is applicable to large scale efforts, but only involving objects which are text documents (i.e. big grain objects), and suffers both of language specificity (multilingual ontologies ca be seen as a step forward towards the largest general ontology) and well-known linguistic troubles (polysemy, synonymy). Most of all, following this approach the meaning must be found all in the words. There is no room for the meaning lying in the use of resources.

The engineers' approach, on the other hand, although it allows for richly and strictly formalized ontologies (e.g. Description Logic ontologies), actually cover just small domains and are suitable for closed environments, where trained people is able to properly classify resources. Here resources nevertheless can be objects of any dimension (data, multimedia resources ...) not only text documents).

Social tagging, finally, shows no interest in producing a sound classification of resources, yet is able to convey communicational intention, a sort of proposal of use of the resources. Thus it provides a dirty-but-working cataloguing of large amounts of data thanks to large communities of users with no specific skills nor training. Most of all, we note that a tag carries both lexical and pragmatic meaning (see figure 1).

Folksonomies indeed are the result of personal free tagging that users do for their own sake, so as to take note of some resource for a possible future use or reference (Vander Wal, 2004). Nevertheless, tags placed by a user can be used also by another user to find and access resources, whenever s/he finds meaningful the association of a given concept (the one referred to by the written content of a tag) to a resource that s/he is actually looking for. Although the word used on such a label may not be the fittest or the most suitable one with respect to the very nature of that resource, the tag puts in evidence some characteristic of the resource that directly calls to some need (or use) for which the resource can be an interesting answer. In this sense it is meaningful to look for different users with compatible points of view on the same set of resources, whihch emerge as possibly overlapping tags assignments.

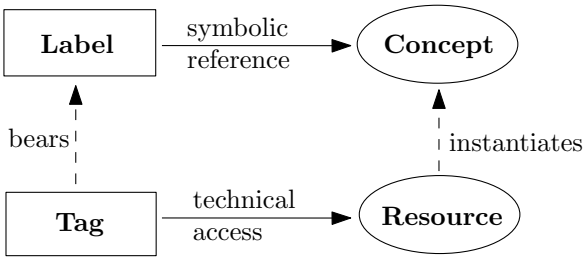

Figure 1: The double nature of tags.

3 A common logical framework

Since the technical access relation (remember of figure 1) works always smoothly, whereas troubles appear as soon as one gets at the lexical level, where labels are supposed to have some conceptual meaning, our idea is to make an attempt to discard this upper level for a while and keep just a minimal logical framework to refer to as the 'meaning' (as value of use) of a resource.

Definition 1 (Compatibility) *We say two resources r and r' are compatible iff the same user u has tagged them with the same tag t.*

We note their compatibility $r \supset r'$. As such a relationship, Compatibility is
simmetric: if $r \supset r'$ then $r' \supset r$ too,
reflexive: each resource is compatible with itself,
not transitive: there may be x, y, z s.t. $x \supset y$, $y \supset z$ but x is not compatible with z – just think of pairs with different tags: $\langle t, x\rangle$, $\langle t, y\rangle$, $\langle t', y\rangle$ and $\langle t', z\rangle$.
Let's now define the space where to look for this compatibility.

Definition 2 (Tagging space) *A tagging space P is the collection of all the pairs $\langle t, r\rangle$ in $P \subseteq T \times R$, that is the set of all the assignments of tags ($t \in T$) to resources ($r \in R$) recorded by one single user u in a tagging community.*

This is to ensure the same intentional meaning behind the use of each single tag.
And finally our logical representation of folksonomies[1]:

[1] As well as ontologies, since behind the rules and fixed conventions followed by trained experts in classifying resources one can see one fixed notion of compatibility.

Definition 3 (Compatibility Space) *A Compatibility Space (CS) X is the web whose support $|X|$ is made of all the resources occurring in the assignments recorded in a given tagging space P, and the points in it are connected according to the Compatibility relationship above defined.*

That is, once fixed the dimension of User by recurring to the notion of tagging space, let $x \in |X|$ and $t \in T$ be respectively

- a resource x appearing in at least one pair in the collection of tag assignments P
- a tag t from the set of tags T used by the fixed user u
- and let $\langle t, x\rangle$ be a recorded tag assignment

so that a Compatibility Space X is more formally defined by its

support: the underlying set of resources, noted $|X|$

compatibility: a binary, reflexive, symmetric, not transitive relation between points of $|X|$, noted $x \frown_X y$, assigned thus: for $x, y \in |X|$

$$x \frown_X y \Leftrightarrow \exists t \in T \text{ s.t. } \langle t, x\rangle \in P \ \wedge \ \langle t, y\rangle \in P$$

As presented in (Abrusci et al., 2011), Compatibility Spaces are a special kind of Girard's Coherence Spaces (Girard et al., 1989). Coherence Spaces provide denotational semantics for Linear Logic (LL) (Girard, 1987): A Coherence Space interprets some formula, whereas operations between Coherence Spaces interpret compositions of formulas according to LL connectives. In particular, within a Coherence Space a subset a of $|X|$ whose points are all pairwise coherent is called *clique*, and is noted $a \sqsubset X$. The notion of *maximal clique* then is what actually in the Space denotes a formula. The point is now: What the Compatibility Spaces interpret? They interpret some notion of compatibility between resources – rough approximation of concepts – by means of the same idea of maximal clique, though with surprising results.

Definition 4 (Clique in a CS) *As for Coherence Spaces, a group a of pairwise compatible points of X is called a* clique*, and is noted $a \sqsubset X$. More formally:*

$$a \sqsubset X \Leftrightarrow a \subset |X| \wedge \forall (x, y) \in a, \ x \frown y$$

But with Compatibility Spaces and the dynamics of tagging we have now three 'flavours' of maximal cliques, corresponding to three different ways of approximating concepts based on three different (stricter of looser) notions of compatibility:

- $\forall x \in |X| \ (\forall y \in a \ \ \exists! t \in T \text{ s.t. } \langle t, x\rangle \in P \wedge \langle t, y\rangle \in P) \ \Rightarrow \ x \in a$ is the strictest notion of compatibility, that gathers only the resources bearing exactly the same tag;
- $\forall x \in |X| \ (\forall y \in a \ \ \exists t \in T \text{ s.t. } \langle t, x\rangle \in P \wedge \langle t, y\rangle \in P) \ \Rightarrow \ x \in a$ is a looser notion of compatibility, allowing for compatibility being limited to any pair of resources occurring in the clique;
- $\forall x \in |X| \ (\forall y \in a \ \ \forall t \in T \text{ s.t. } \langle t, x\rangle \in P \wedge \langle t, y\rangle \in P) \ \Rightarrow \ x \in a$ finally is quite a strange notion of compatibility looking for, so to say, maximally compatible resources, i.e. a clique collecting only resources that all share all their tags.

4 Perspectives

Whereas from a theoretical point of view quite an important work is yet needed to have the notion of Compatibility Space perfectly fit within the scenery of Linear Logic and its subsequent developments such as Ludics[2] (Girard, 2001), the most interesting part will be now to develop some test cases to assess the actual usefulness of such an interpretation of ontologies and folksonomies by determining the relevance and the cognitive value of the concepts approximated by maximal cliques, possibly also considering the different kinds of folksonomies that Vander Wal (2005) distinguishes.

References

V.M. Abrusci, M. Romano and C. Fouqueré. 2011. Ontologies and Coherence Spaces. In *Ludics, Dialogue and Interaction.* Volume 6505 of *FoLLI LNAI*, pp. 205-219. Springer-Verlag.

J.-Y. Girard. 1987. Linear Logic. In *Theoretical Computer Science.* Volume 50, pp. 1-102.

J.-Y. Girard, Y. Lafont and P. Taylor. 1989. *Proofs and Types.* Cambridge University Press.

J.-Y. Girard. 2001. Locus solum: From the rules of logic to the logic of rules. In *Math. Structures in Computer Science.* Volume 11(3), pp. 301-506.

M. Romano. 2011. *Ontologies, Logic and Interaction.* PhD Thesis. `http://logica.uniroma3.it/files/MRomanoPhDThesis.pdf`

T. Vander Wal. 2004. *Folksonomy.* `http://vanderwal.net/folksonomy.html`

T. Vander Wal. 2005. *Explaining and Showing Broad and Narrow Folksonomies.* `http://vanderwal.net/folksonomy.html`

[2](Romano, 2011) introduces the basic lines of a sort of querying protocol over Web datasources presented as Compatibility Spaces.

Towards an ontology based information system for Linguistic: the case study of the OTIM project

First author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

1 Introduction

This work stands in the OTIM (Tools for Multimodal Annotation processing) project [1]. It aims at developing conventions and tools for multimodal annotation of a large conversational French speech corpus (Blache et al., 2010) (http://aune.lpl.univ-aix.fr/~otim/). OTIM can be summarized in two main steps.

The first step concerns *the multimodal annotation of a conversational speech between two persons*. It is under the responsibility of linguists; annotation is done according to different levels of linguistic analysis. Each expert has to annotate the same data flow according to its knowledge domain and the nature of the signal on which he annotates (signal transcription or signal). Experts generally use dedicated tools like PRAAT[2], ANVIL[3] or ELAN[4]. The qualifier multimodal is due to the nature of the studied corpus which is composed of text, sound, video. Within the project OTIM, linguists propose an encoding for annotating spoken language data, with the acoustic signal as well as its orthographic transcription. They have chosen to use Typed Feature Structures (Carpenter, 1992)(Copestake, 2003) (TFS) to represent in an unified view the knowledge and the information they need for annotation. Linguistic annotation tools rely on native and not often open formats which are not directly interoperable. TFS provides an abstract description using a high level formalism independent from coding languages and tools.

The second step concerns *the representation and manipulation of multimodal annotation*. We aim at providing linguists with a unique framework to encode and manipulate numerous linguistic domains (morpho-syntax, prosody, phonetics, disfluencies, discourse, gesture and posture (Blache et al., 2010)) in order to analyze and find correlations between annotated linguistic domains. For that, it has to be possible to bring together and align all the different annotations associated to a corpus.

In this paper, we focus on this last step considering semantic web technologies for the development of a Knowledge-based Information System.

2 Context and Motivation

Linguistic knowledge is captured by means of three types of information : *properties* (the set of characteristics of an object); *relations* (the set of relations that an object has with other objects); *constituents* (complex objects composed of other objects). TFS proposes a formal presentation of each annotation in terms of feature structures and type hierarchies : properties are encoded by features, constituency is implemented with complex features, and relations make use feature structure indexing; each linguistic domain is represented as a hierarchical model. TFS enables linguists to represent in an unified view the knowledge and the information they need for annotation.

Figure 1 graphically describes TFS representation of the prosodic domain; a formal definition can be found in (Carpenter, 1992) (Copestake, 2003). For sake of simplicity, we do not detail the meaning of every feature used in the example.

Due to its theoretical nature, TFS representation cannot be used within an applicative framework

[1] supported by the French ANR agency
[2] http://www.fon.hum.uva.nl/praat/
[3] http://www.anvil-software.de/
[4] http://www.lat-mpi.eu/tools/elan/

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 22–24
Paris, 10 November 2011

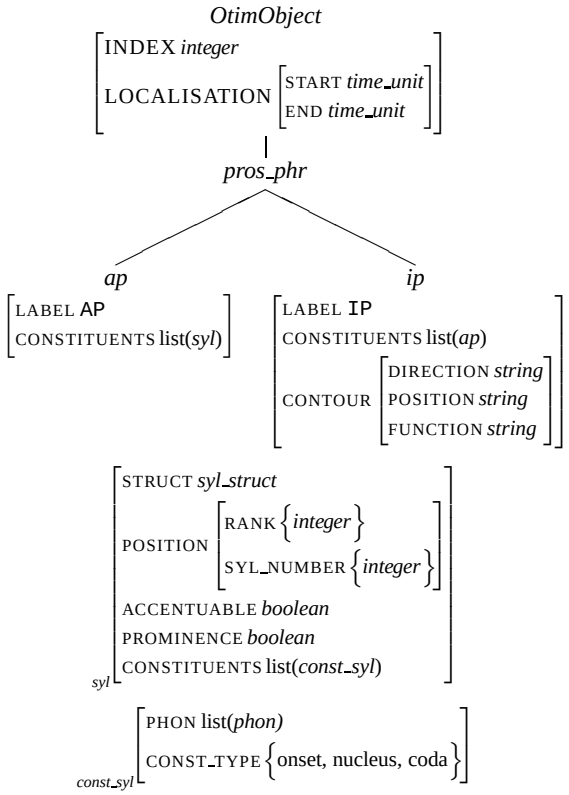

Figure 1: TFS representation of the prosodic domain

and has to be implemented into other formalisms.

3 Contributions

We have propose an knowledge representation formalism which be an alternative to TFS : an ontological approach based on Description Logics (Baader et al., 2003) (DL) and on semantic web technologies for the development of a linguistic Knowledge-based Information System(Seinturier, 2011).

Some linguistic projects have a similar objective than OTIM, for instance NITE[5], AGTK[6], PAULA[7], XStandoff (Stuhrenberg, 2009). Our approach differs from them because we focus on an ontological contribution. These proposals generally propose toolkits for multi-level annotation by means of libraries of data and annotation management. Moreover, linguistic annotation tools rely on native and not often open formats which are not directly interoperable. The multiplication of annotation schemes and coding formats is a severe limitation for interoperability. One solution consists in developing higher level approaches (Ide, 2007)(Stuhrenberg, 2009). However, these experiments still remain very programmatic.

3.1 Creating OWL ontology

Creation of the OWL ontology follows two steps. First of all, the terminological knowledge from the TFS is implemented into OWL using the Protege[8] ontology editor. The Protege framework was initially designed for biologists and biochemists. This characteristic is quite interesting because this is not a computer scientist tool and so there is no need of a specific knowledge in computer science to use it.

Figure 2 shows the ontology of the prosodic domain. This ontology is linked with two other domains: the phonetics domain, which is a part of the OTIM knowledge representation framework, and the time domain given by a standard ontology of the W3C.

3.2 Managing data and querying with SPARQL

Management and querying of OWL data relies on the standard SPARQL (Prudhommeaux, 2007) querying language. SPARQL enables to match graph pattern against the graph of RDF/OWL triple ($WHERE$ clause) and identifies values to be returned ($SELECT$ clause). The $FROM$ clause enables to identify the data sources to query.

We express within the OTIM project the linguistic inter domain queries designed on TFS by SPARQL queries on the OWL representation. A sample query expressed in natural language is:

"We need the list of phonemes that are associated with the accentual phrases stated between the second 35 and the second 55 of the speech."

This query takes into account the prosodic domain (accentual phrase), the phonetic domain (phoneme) and the time. Such a query is represented in SPARQL by:

```
1.  SELECT      ?phoneme
2.  FROM        otim − prosody.owl, otim − phonetics.owl
3.  WHERE {     ?const rdf:type prosody:SyllableConst
4.              . ?const hasPhonemes ?phoneme
5.              . ?syl rdf:type prosody:Syllable
6.              . ?sc hasConstituents ?const
7.              . ?ap rdf:type prosody:AccentualPhrase
8.              . ?ap hasSyllables ?syl
9.              . ?t rdf:type time:TemporalEntity
10.             . ?ap hasTimeLocation ?t
11.             . ?tref time:contains ?t }
```

[5]http://groups.inf.ed.ac.uk/nxt/

[6]http://weblex.ens-lsh.fr/projects/xitools/logiciels/AGTK/agtk.htm

[7]http://www.sfb632.uni-potsdam.de/ d1/paula/doc/

[8]http://protege.stanford.edu/

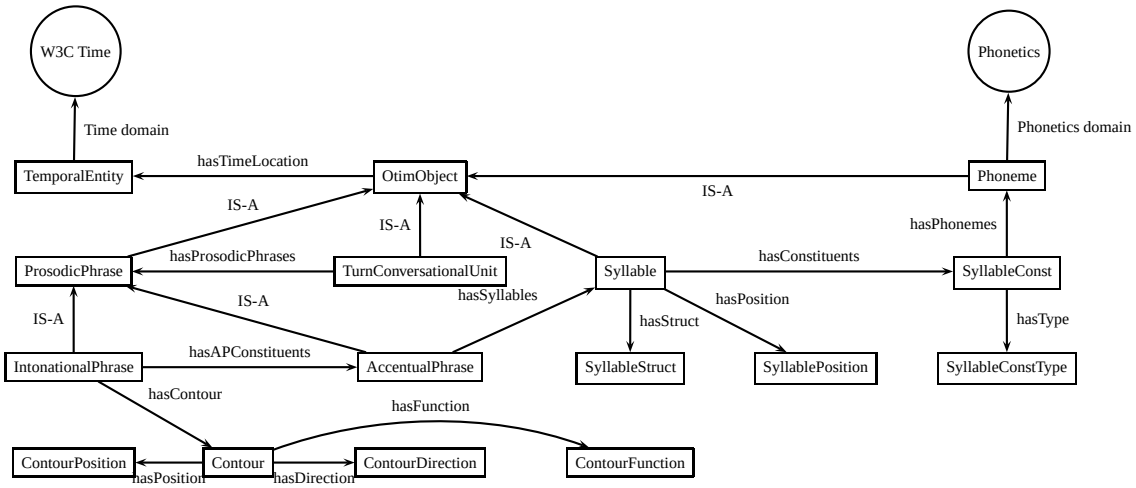

Figure 2: Ontological representation of the prosodic domain

We assume ithat the time bounds given are represented as a $TemporalEntity$ named $tref$. The $SELECT$ clause specifies that the result to build is made of phonemes. The clause $FROM$ contains the two data sources on which the query is processed (the two target domains prosody and phonetics). The $WHERE$ clause describes the patterns for a phoneme to match.

The OTIM framework for linguistic multimodal annotations management has been implanted within a Java/OWL framework. The OWL standard used is OWL-DL as this is the specification that gives all the expressiveness we need and guarantees some calculability results that are critical for querying data.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

J. Allen. 1991. Time and time again : The many ways to represent time. *International Journal of Intelligent Systems*, 6(4):341–355, july.

Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

Philippe Blache, Roxane Bertrand, Brigitte Bigi, Emmanuel Bruno, E Cela, Robert Esperrer, Gaelle Ferre, Marion Guardiola, Daniel Hirst, Ep Magro, Martin JC, Meunier C, Morel Ma, Elisabeth Murisasco, Nesterenko I, Nocera P, Pallaud B, Prevot Laurent, Priego-Valverde J, Julien Seinturier, Tan N, Tellier Marion, and Rauzy Stephane. 2010. Multimodal annotation of conversational data. In *Proceedings of the fourth linguistic annotation workwhop (LAW)*, pages 186–191. Assocation for computational Lunguistics (ACL), 15 july 2010.

Robert L. Carpenter. 1992. *The Logic of Typed Feature Structures*, volume 32 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, The Edinburgh Building, Shaftesbury Road, Cambridge CB2 8RU, United Kingdom, c. j. van rijsbergen edition, 1992.

Ann Copestake. 2003. *Collaborative Language Engineering: A Case Study in Efficient Grammar-based Processing*, chapter Definitions of Typed Feature Structures. CSLI Publications, Ventura Hall, Stanford University, Stanford, CA 94305-4115, 2003.

N. Ide and K. Suderman. 2007. Graf: A graph-based format for linguistic annotations. In *Linguistic annotation workshop (LAW)*. Association for computational Lunguistics (ACL), 2007.

Eric Prud'hommeaux and Andy Seaborne. 2007. Sparql query language for rdf (working draft). Technical report, W3C, March 2007.

J. Seinturier, H. Rouine, E. Murisasco, and E. Bruno andP. Blache. 2011. Knowledge-based multimodal data represnetation and querying. In *Int. conf. on Knowledge Engineering and Ontology Developmement (KEOD 2011)*, Paris, France, October 2011.

M. Stührenberg and D. Jettka. 2009. A toolkit for multi-dimensional markup - the development of sgf to xstandoff. In *Proceedings of Balisage: The Markup Conference 2009*. Assocation for computational Lunguistics (ACL), Balisage Series on Markup Technologies, vol. 3, 2009.

On the ontological coherence of lexical resources

Laure Vieu
IRIT – CNRS – Université de Toulouse
and LOA – ISTC – CNR
vieu@irit.fr

1 The need for methods

Lexical resources like WordNet are increasingly exploited in NLP tasks reasoning on their structure –especially their taxonomy–, for instance, semantic similarity or recognizing textual entailment. Such lexical resources are even used more generally as ontologies for their large coverage. They nevertheless are *not* ontologies, and it is well known in particular that WordNet's structure is riddled with ontological errors (Kaplan and Schubert, 2001; Gangemi et al., 2003; Clark et al., 2006), affecting the quality of the tasks in which it is exploited (Neel and Garzon, 2010).

Any effort to build an "ontological lexicon", i.e., a lexical resource with a significant ontological structure, should first invest in a methodology to guarantee its ontological coherence. Top-level ontologies are small enough to be hand-crafted and fully checked by formal ontologists (although this is not always the case). On the other hand, lexical resources are both typically much larger and likely to be built by lexicographers and terminologists rather than formal ontologists, or even to be semi-automatically built from other resources and data.

Existing proposals, be they coherence verification methods like the rigorous but difficult OntoClean (Guarino and Welty, 2004) and the simpler (Alvez et al., 2008), or lexicon-ontology alignments (Pease and Fellbaum, 2009), are all manual, thus prone to error and very costly to implement. I report here on the feasibility of semi-automatic methods to "clean" existing lexical resources or to assist the construction of new ones. I have shown in (Verdezoto and Vieu, 2011) on examples from WordNet that it is possible to detect several types of errors automatically. Recurring problems clearly appear in this experiment. Their analysis allows the writing of guidelines to support the manual fixing of the errors so detected, thus completing the semi-automatic method looked for.

2 Automatic detection of incoherences

The idea is to contrast two sources of knowledge involving two different ontological relations, e.g., the taxonomic relation ISA and the (or *one*) meronymic relation Part-of. On the basis of semantic and ontological constraints inherent to these relations, we can then automatically detect local incoherences between these two sources.

This requires that the two sources of knowledge are already aligned. This is naturally the case *within* rich lexical resources that do have more than one structural relation, such as WordNet or FrameNet. Another avenue (of course not available during the construction of the resource) is to contrast the structure of the lexical resource with corpora annotated with another[1] relation, provided the corpora are also annotated with word senses of the lexical resource itself, as in some data sets of the SemEval benchmark (Girju et al., 2007).

3 Semantics of structural relations

The first ontological relation present in lexical resources arguably is the taxonomic ISA. When discussed at all, its semantics is generally admitted to be class inclusion. Similarly, Instance-of is membership of an individual in a class.[2]

[1] With a same relation, the coherence check is direct.

[2] Instances appear in lexical resources when proper names are covered.

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 25–27
Paris, 10 November 2011

The semantics of other relations is much less clear. Regarding the meronymic Part-of, there is at most a distinction between *several* relations of Part-of (e.g., "member" Part-of and "component" Part-of, usually along the lines of (Winston et al., 1987)). Although clarifying as much as possible the semantics in formal terms is highly desirable, this is by no means straightforward, especially since relations like meronymic ones in lexical resources often have a typical flavour[3] that is still foreign to standard ontology theories and methods. Fortunately, it appears that fully specifying the semantics is not strictly necessary to start exploiting some coherence constraints, as simple questions can be focussed on and answered.

4 Constraints between relations

When contrasting taxonomic ISA and Instance-of with other relations (e.g., meronymic), two main questions appear, already providing useful constraints.

First, how is the individual / class distinction affecting the other relations? That is, do these relations equally make sense when applied to any combination of arguments, individual-individual, class-class, individual-class and class-individual? In the case of Part-of, it turns out that the individual-class option is to be excluded, yielding a constraint to be checked for.

Second, are there any categorial constraints on the arguments of the relation? Should one argument be of a given top-level category? For instance, the specific "member" Part-of relation requires its second argument (the whole) to be a collection, another constraint to be checked. Should the categories of the two arguments be related in a specific way? For instance, the Part-of relation imposes a categorial homogeneity between the part and the whole. Suppose we adopt a standard top-level distinction between concrete objects, events and abstract entities. Then, these three disjoint categories cannot be mixed in a Part-of relation,[4] thus yielding yet other constraints.

[3]For example, the relationship between *handle* and *hand tool* is not necessary for neither handles, which also appear as parts of doors and briefcases, nor for hand tools, which include handleless scrapers.

[4]For example, your ride to work can have sub-events as parts, but your bike is not a part of it.

Automatically checking whether such constraints are respected and spotting the corresponding errors is then straightforward. Resolving them is less so, but many regularities do appear among them.

5 What errors tell us

A local incoherence between two sources of knowledge is a symptom of an error that may appear in one source or the other. In our example, the error is either in the taxonomic relations or in the meronymic ones.

In the experiment made on WordNet in (Verdezoto and Vieu, 2011), meronymic errors mostly point at confusions on the Part-of relations with relations like participation, indicating as expected that it is necessary to clarify their semantics. Taxonomic errors are particularly telling, since their semantics is in principle clearer. Classical ISA overloading (confusion between ISA and Instance-of, confusion between class and collection) shows up and it is relatively easy to help a user solving it. More interestingly, errors related to lexical polysemy occur. In particular, systematic polysemy of the kind accounted for with "dot objects" in (Pustejovsky, 1995), like with *book* (bunch of paper sheets and text), has not been uniformly addressed in WordNet and the solution of multiple inheritance adopted in places yields ontological clashes (nothing can be both concrete and abstract). Another surprising issue for a lexical resource with a very refined notion of word sense is that multiple senses are not everywhere coherently inherited throughout the taxonomic hierarchy.[5]

As a conclusion, the work reported here proves the feasibility of methods based on the automatic detection of incoherences in lexical resources. In addition, it shows that such methods can reveal structural design shortcomings in lexical resources, as with WordNet's missing schemes to handle systematic polysemy or to enforce multiple sense inheritance. These of course should be solved before any complete method aimed at semi-automatically ontologically cleaning the lexical resource can be implemented.

[5]E.g., *country* has both the sense of "location" and that of "people", but a specific country like *Ethiopia* has a unique sense which appears as an instance of "location", the "people" sense being missing.

References

Javier Alvez, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver, and German Rigau. 2008. Complete and consistent annotation of WordNet using the Top Concept Ontology. In *Proceedings of LREC2008*, pages 1529–1534.

Peter Clark, Phil Harrison, Tom Jenkins, John Thompson, and Rick Wojcik. 2006. From WordNet to a Knowlege Base. In Chitta Baral, editor, *Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering. Papers from the 2006 AAAI Spring Symposium*, pages 10–15. AAAI Press.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003. Sweetening WordNet with DOLCE. *AI Magazine*, 24(3):13–24.

Roxana Girju, Vivi Nastase, and Peter Turney. 2007. SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18. Association for Computational Linguistics.

Nicola Guarino and Chris Welty. 2004. An overview of OntoClean. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 151–159. Springer-Verlag.

Aaron N. Kaplan and Lenhart K. Schubert. 2001. Measuring and Improving the Quality of World Knowledge Extracted From WordNet. Technical Report 751, University of Rochester.

Andrew Neel and Max Garzon. 2010. Semantic Methods for Textual Entailment: How Much World Knowledge is Enough? In *Proceedings of FLAIRS 2010*, pages 253–258.

Adam Pease and Christiane Fellbaum. 2009. Formal ontology as interlingua: the SUMO and WordNet linking project and Global WordNet. In Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari, and Laurent Prévot, editors, *Ontology and the Lexicon. A Natural Language Processing Perspective*, pages 31–45. Cambridge University Press.

James Pustejovsky. 1995. *The generative lexicon*. MIT Press, Cambridge (MA).

Nervo Verdezoto and Laure Vieu. 2011. Towards semi-automatic methods for improving WordNet. In Johan Bos and Stephen Pulman, editors, *Proceedings of the Ninth International Conference on Computational Semantics IWCS 2011*, pages 275–284, Oxford, UK.

Morton Winston, Roger Chaffin, and Douglas Herrmann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.

An Ontological and Terminological Meta-model for Semantic Information Retrieval

Axel Reymonet
ACTIA Research Team
25 Chemin de Pouvourville
31432 Toulouse, FRANCE
axel.reymonet@actia.fr

Jérôme Thomas
ACTIA Research Team
25 Chemin de Pouvourville
31432 Toulouse, FRANCE
jerome.thomas@actia.fr

Nathalie Aussenac-Gilles
IRIT - Melodi, UPS
118 Route de Narbonne
31062 Toulouse, FRANCE
nathalie.aussenac@irit.fr

1 Context and objectives

Based on statistical techniques applied to the lexical content of documents along with the number and popularity of referrers (Baeza-Yates and Ribeiro-Neto, 1999), traditional search engines do not allow a semantically enhanced description of their search base. This leads to a lack of precision in the results, caused by the ignorance of synonymy and the possible ambiguities of the requests, as stated in (Berners-Lee, 1999).

Semantic Information Retrieval (IR) on a specialized domain copes with these issues by connecting the traditional weighted strings to the ideas they bear. Working on a conceptual level logically requires a prior model of the knowledge found in the technical documents, hence a common framework to handle both notions of ontology and terminology: the Ontological and Terminological Resource (OTR).

2 A meta-model for Semantic Indexing

In the last 5 years, Knowledge Engineering from texts has emerged as a promising way to reduce the cost for building and maintaining domain ontologies (Buitelaar et al., 2005). Such an ontology learning process makes it possible to keep tracks of modelling choices and to connect many lexical entries (or terms) to concepts. This feature is all the more relevant when the ontology is used to associate a document with a formal interpretation of (part of) its textual content.

2.1 First version

After a thorough analysis in (Reymonet et al., 2007), we came to the conclusion that the existing models allowing joint representation of an ontology and its associated terminology were insufficient: firstly, the notion of term is barely reified in the literature, whereas it could help to model any information related to the linguistic manifestation of a term. Moreover, most classical models (e.g. those based on GATE, described in (Bontcheva et al., 2004)) connect a concept and a term by means of a class-instance relation, which does not allow a faithful reproduction of some linguistic phenomenons such as anaphora or polysemy. Willing to respect the most popular ontological standard, we proposed in (Reymonet et al., 2007) a first meta-model compatible with OWL-DL to store terms without altering the syntax of the sublanguage. However, some operational choices made our model theoretically flawed. That is why we decided to bring some changes to it by resorting to OWL-Full meta-modelling primitives. In (Reymonet et al., 2009), we converged to the solution depicted below.

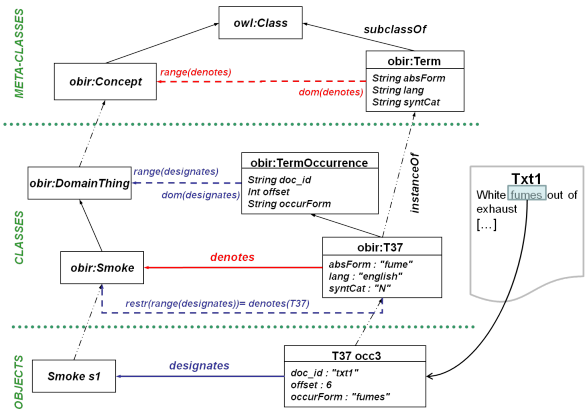

Figure 1: Our first "solid" OTR meta-model

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 28–29
Paris, 10 November 2011

2.2 Enriched version

The meta-model illustrated in figure 1 enabled us to manipulate both terms and concepts independently during the construction and maintenance of the OTR. However, considering the ability to use automatic reasoners on OWL ontologies, we decided to enrich further the current meta-model, in order to make it more suitable for a semantic indexing task. As a matter of fact, reifying document annotations (see fig. 2) grants us the ability to find quickly (e.g. using a SPARQL query) all documents whose annotations do not respect some basic rules: no inconsistency, a minimum ratio of found words...

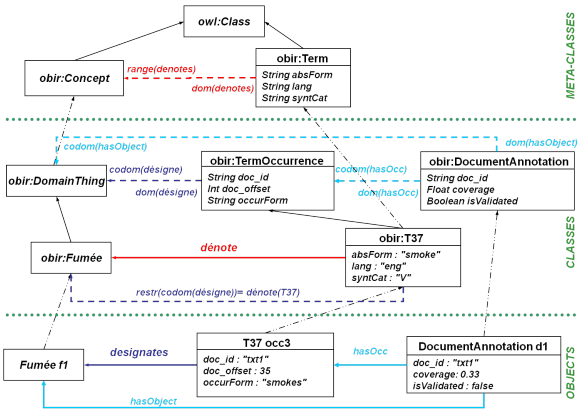

Figure 2: Partial example on enriched meta-model

3 An application to automotive diagnosis

Thanks to both the Lucene API [1] and our OTR meta-model specifically designed for semantic indexing, we were able to build a Protégé-OWL [2] plugin covering both aspects of semantic IR, i.e. indexing and searching. We mainly use it on a troubleshooting base in automotive diagnosis, but our plugin is completely generic, hence able to run on ontologies and data from very different domains.

The authoring tool allows to carry out in parallel the OTR completion/maintenance and the semantic indexing of the corpus. The texts are automatically indexed by the lexicalizations already present in the OTR and the users are presented with the results. If the document annotation is unsatisfying, they can react by correcting the semantic indexes and/or modifying the OTR accordingly. By carrying out two tasks at the same time, the plugin is aimed both at reducing time manipulating the OTR and at relieving the modeller from the burden of achieving manually its construction.

Once the indexing process is finished, the runtime tool gives access to a semantic search engine. It automatically transforms any free-text request into its (sometimes partial) semantic representation in the given OTR. The engine then computes a semantic proximity between the query and each document. Several state-of-the-art conceptual similarities such as (Wu and Palmer, 1994; Lin, 1998) are used on that purpose. In the end, only the documents whose proximity with the request is higher than a given threshold are shown to the user, ordered by decreasing similarity.

[1] http://lucene.apache.org/java/docs/

[2] http://protege.stanford.edu/overview/protege-owl.html

References

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999. ISBN 020139829X.

Tim Berners-Lee. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper San Francisco, 1999. ISBN 1402842937.

K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10(3/4):349–373, 2004.

P. Buitelaar, P. Cimiano, and B. Magnini, editors. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, 2005.

Dekang Lin. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

Axel Reymonet, Jerome Thomas, and Nathalie Aussenac-Gilles. Modelling ontological and terminological resources in OWL-DL. In *Proceedings of ISWC '07 workshop "From Text to Knowledge: The Lexicon/Ontology Interface" (OntoLex '07)*, November 2007.

Axel Reymonet, Jerome Thomas, and Nathalie Aussenac-Gilles. Ontology based Information Retrieval: an application to automotive diagnosis. In *Proceedings of 2009 International Workshop on Principles of Diagnosis (DX'09)*, June 2009.

Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133–138, New Mexico State University, Las Cruces, New Mexico, 1994.

Text-based IE and Open Linguistic Data for Termontological Resources

Andrés Domínguez Burgos **Koen Kerremans** **Rita Temmerman**

Centre for Special Language Studies and Communication
Erasmus University College Brussels
Pleinlaan 5
B-1050 Brussels
Belgium

adres.dominguez.burgos@ehb.be koen.kerremans@ehb.be rita.temmerman@ehb.be

Abstract

We present work in progress concerning the development of a multilingual termontological resource dealing with cultural events. The article will describe how the termontological resource can benefit from text-based information extraction (IE) and open linguistic data.

Keywords. computational terminology, text-based information extraction (IE), linked data, multilingual termbase, open linguistic data, open linguistics, termontography

1 Introduction

The work presented in this article is part of the project *Open Semantic Cloud for Brussels* (OSCB)[1]. This project aims to cover Brussels with a cloud of structured and interlinked information elements produced by "atomizing" a collection of relevant databases and other resources. The meaningful exploitation of such interlinked cloud-type resources requires the adoption of ontologies. These ontologies can be used for automatically annotating information in databases on the Web.
Our objective is to show how in the framework of the OSCB project, a multilingual (French, Dutch, English) termontological resource will be developed that will need to support the development of the OSCB ontology. This is why – unlike many traditional terminology resources that only provide information about the linguistic properties of terms and their syntactic relations – the resource needs to provide both linguistic as well as conceptual information (Temmerman & Kerremans 2003).

In this article, we present part of the data model of the termontological resource (section 2). Next, on the basis of a description of some state-of-the-art work in computational terminology and open linguistics (section 3), two main issues will be addressed. First, how can recent trends in the automatic extraction of terminological data boost the development of the multilingual termontological resource? Second, how can the termontological resource benefit from initiatives related to open linguistic data? These two questions will be addressed in section 4.

2 The termontological data model

Part of the data model of the termontological resource is shown in the figure below. The data categories in bold were taken from a global dataset managed by TDG 9, the Thematic Domain Group on Terminology, in the context of ISO 1260:2009[2].

- conceptual level
  - **conceptIdentifier**
    - concept description
    - **conceptRelation**
      - **associatedConcept**
      - **genericRelation**
      - …
    - …
  - **ontologyConceptEntry**
  - **subjectField**
- language level
  - term
    - **termType**
    - term property
    - …
- …

Figure 1. Sample of the termontological data model

At the conceptual level, we distinguish between the 'conceptIdentifier (i.e. a unique label for the terminological record), the 'ontologyConceptEntry'

[1] This project is financed by Innoviris, the Brussels Institute for Research and Innovation (http://www.innoviris.be/). More information about OSCB can be found here: http://www.oscb.be

[2] For more information: http://www.isocat.org/files/TDGs.html

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 30–32
Paris, 10 November 2011

(i.e. a label to indicate the corresponding concept in the OSCB ontology) and the 'subjectField' (i.e. an indication of the subject field(s)). The 'conceptIdentifier' features a concept description and a set of conceptual relations.
At the language level, the termontological resource provides information about the term, such as the 'termType' (e.g. full form or short form) or term properties (e.g. POS tag, grammatical gender).

3 Text-based IE and open linguistic data

The OSCB multilingual termontological resource needs to be 1) *flexible* (i.e. adaptable to specific user needs); 2) *dynamic* (i.e. capable of keeping up with possible conceptual and linguistic changes in the subject field); and 3) *useful for both humans and machines* (i.e. linguistic and conceptual information needs to be encoded explicitly) .
One way to create flexible and dynamic termontological resources for humans and machines is to apply information extraction techniques to a multilingual domain-specific monitor corpus – i.e. an ever-growing collection of multilingual texts related to a specific subject area (section 3.1). A second way is by linking terminological data on the web (section 3.2).

3.1 Text-based IE

In *knowledge-based approaches*, systems rely on lists of linguistic surface patterns in order to carry out extraction tasks (Halskov & Barrière 2008). These approaches have the advantage of reaching high precision in certain domains but suffer from scalability. Developing datasets of linguistic patterns requires a lot of time and human effort and the results are not always transferable from one domain to the other. Moreover, they are language-specific.
In *machine learning approaches*, systems are trained to extract different types of terminological information or cluster terms based on a set of features. Different from knowledge-based approaches, is the fact that all machine learning approaches derive knowledge from data, rather than from human analysts. The approaches vary as to the nature of the training material, the degree of human intervention, the types of knowledge required, etc. (Cimiano & Staab 2005; Andrews & Ramakrishnan 2008; Kozareva & Hovy 2010).
The extraction of linguistic and conceptual information requires strategies for improving sense disambiguation. A large range of methods have been developed to deal with solving sense disambiguation (Navigli 2009 for an overview). Navigli and Velardi (2004), for instance, used the so-called structural semantic interconnections approach (SSI) approach. Departing from WordNet synsets, they produced graphs representing possible candidates for the senses of a term and assign a probability to each possible sense according to the context terms.

3.2 Open linguistic data

The idea of using the web for exposing, sharing and connecting pieces of data and information, is getting more attention in the field of linguistics. Today, many types of (freely available) language resources – e.g. annotated text corpora, general language lexicons, thesauri, plain wordlists, specialised glossaries, frequency dictionaries, translation memories, etc. – are scattered throughout the web, waiting to be linked and deployed for specific purposes or applications. Existing frameworks for modelling and representing lexical resources are not applicable to the new type of lexical resource resulting from different linking operations. For example, while the ISO Lexical Markup Framework or ISO 24613:2008 provides useful constructs to represent a range of lexicons, it still concentrates on modelling one lexical resource at a time, and does not provide effective devices to integrate different types of lexical resources into a single combined resource (Hayashi 2011).
Several initiatives have recently been taken to publish and make available different types of linguistic resources as Linked Data, such as the Lemon RDF model (McCrae et al. 2010). The number of linguistic projects involved in Linked Data have grown to such an extent that the need becomes more important to share ideas on how these linguistic data should be made available (e.g. in which formats or via which platforms), how they can be linked or how they can be used to collaborate and share work.

4 Developing termontological resources

Figure 2 illustrates part of the method that we have worked out for structuring multilingual terminology in a categorisation framework, i.e. a conceptual structure showing the important conceptual categories within a specialised domain (Temmerman and Kerremans 2003). The figure also shows how this method can be used for multilingual and semantic information extraction. In an initial step, we apply statistical and linguistic methods to extract lists of seed terms from multilingual specialised corpora. Next, the extracted terminology list is structured in a categorisation framework or CF, showing the link between a term and a category in a structure which is more complex than a taxonomy as it also covers non-taxonomic relations.
The seeds are used to identify relevant documents and, in these documents, new terms. Many of these new terms will be multiword terms. In order to structure these terms within the CF, we will start from WordNet synsets related to possible word fields found

in the relevant domains and use network traversal to determine the corresponding categories. By comparing these categories with categories assigned to Wikipedia articles, the multiword term list will be complemented with multiword terms from Wikipedia that are related to the domain being analysed.

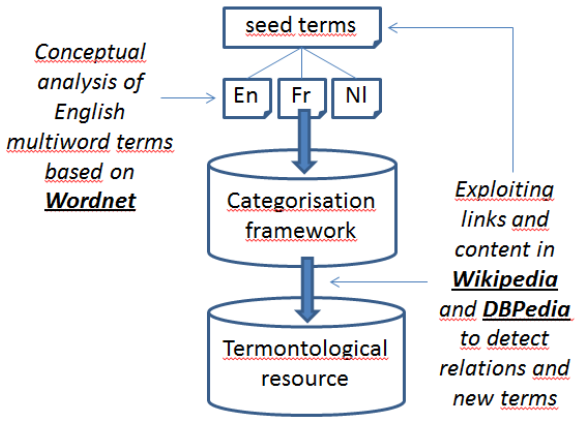

Figure 2. Developing a termontological resource

In our approach, we design a general lexical framework that can represent a wide range of multiword terms as semantic sub-graphs with a set of relationships beyond the basic WordNet relationships. The general meaning of most multiword terms can to a certain extent be determined on the basis of an analysis of their components. We will apply Bayesian inference mechanisms 1) to determine the relations between multiword term components and 2) to be able to link multiword terms in WordNet and Wikipedia to basic categories in the CF. The context in which the multiword terms appear will help us to determine the most likely categories to which they should be attached.

We use Wikipedia – both text and metadata – as main source for mining terminologically and ontologically relevant information in order to 1) expand the search through new Wikipedia articles, 2) determine the scope of relevant terms and possible references to categories in the CF and 3) retrieve domain-specific relationships that could be useful for ontological work. We also use DBPedia RDFs to mine new linguistic patterns and new actual relationships within Wikipedia's free text: most RDFs in DBPedia came from the structured data within Wikipedia entries that also have the same information in unstructured, i.e. free text form. We are designing strategies to associate the RDF data with their most likely corresponding linguistic patterns occurring in free text. This will help us to automate the discovery of new linguistic constructions to express in different languages the types of relationships found in the RDFs.

5 Conclusion

In this article, we have presented work in progress with respect to the development of multilingual termontological resources. We have discussed how termontological resources can benefit from recent trends in text-based IE and open linguistic data. We are currently developing an application supporting the methodology discussed in section 4. The resulting application will allow us to export the termontological resource to the Lemon RDF model so that data can be shared with others (McCrae et al. 2010).

6 References

Andrews, N. & Ramakrishnan, N., 2008. Seeded Discovery of Base Relations in Large Corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 591–599.

Cimiano, P. & Staab, S., 2005. Learning concept hierarchies from text with a guided agglomerative clustering algorithm. *Proceedings of the workshop on learning and extending lexical ontologies with machine learning methods*.

Halskov, J. & Barrière, C., 2008. Web-based extraction of semantic relation instances for terminology work. *Terminology*, 14(1), pp.20-44.

Hayashi, Y., 2011. A Representation Framework for Cross-lingual/Interlingual Lexical Semantic Correspondences. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*. pp. 155-164.

Kozareva, Z. & Hovy, E., 2010. A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 1110–1118.

McCrae, J. et al., 2010. *The lemon cookbook*.

Navigli, R. & Velardi, P., 2004. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30.

Navigli, R, 2009. Word Sense Disambiguation: a Survey. ACM Computing Surveys, 41 (2). ACM Press, p-169.

Temmerman, R. & Kerremans, K. 2003. Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description. In *Proceedings of CIL17*. Proceedings of CIL17. Prague: Matfyzpress.

Ontology Lexicalisation: The *lemon* Perspective

Paul Buitelaar[•], Philipp Cimiano[⋆], John McCrae[⋆], Elena Montiel-Ponsoda[†], Thierry Declerck[‡]
[•]Unit for Natural Language Processing, DERI, National University of Ireland, Galway
[⋆]Semantic Computing Group, CITEC, University of Bielefeld, Germany,
[†]Ontology Engineering Group, Universidad Politécnica de Madrid, Spain,
[‡]Language Technology Lab, DFKI, Germany

1 Introduction

Ontologies (Guarino1998) capture knowledge but fail to capture the structure and use of terms in expressing and referring to this knowledge in natural language. The structure and use of terms is the concern of terminology as well as lexicology. In recent years, the relevance of terminology in knowledge representation has been recognized again (for example the advent of SKOS[1]) but less consideration has been given to lexical and linguistic issues in knowledge representation (Buitelaar2010).

2 Use Cases of Ontology Lexicalisation

Natural language is often the medium of choice for knowledge representation and transfer between humans. However, ambiguity is widespread in natural language. Words have multiple meanings and grammar can be ambiguous in structure and therefore in interpretation. However, such ambiguities appear to provide little issue to people, who can with little effort resolve these ambiguities in nearly all situations. Machines, on the other hand, have significant issues in resolving these ambiguities and this can lead to difficulties in defining precise interpretations in technical domains. To illustrate this we will now briefly explore some of the use cases of ontology lexicalisation, i.e. in knowledge acquisition from text and multilingual knowledge access.

2.1 Knowledge Acquisition from Text

In the case of knowledge acquisition from text we aim to identify relevant text segments and align these with formally defined knowledge structures, such as facts and axioms. Let us focus on ontology-based information extraction, that is, the extraction of facts from text relative to a given ontology. Consider for example an ontology on tourism with ontology labels (terms) in Spanish. The ontology defines concepts of relevance to tourism such as historical buildings, which will be defined by use of the Spanish term (ontology label) "edificio histórico". For instance, in the following sentence there is a specification of a set of facts concerning a historical building (*Universidad de Barcelona*), its architect (*Elies Rogent*), and building period (*1863-1882*):

- "El edificio histórico de la Universidad de Barcelona es obra de Elies Rogent, se inició su construcción en 1863, pero no se concluyó hasta 1882." (The historical building of the University of Barcelona is the work of Elies Rogent, its construction began in 1863, but was not completed until 1882.)

Observe that the match between ontology label and text is straightforward, as they are identical. However, this is not the case in the following example:

- "El Cabildo de Buenos Aires, ... El edificio, declarado Monumento Histórico Nacional desde el año 1933, fue objeto de sucesivas alteraciones, ... " (The Cabildo of Buenos Aires, ... The building, declared a National Historic Landmark in the year 1933, underwent successive alterations, ...)

In this case, the text segment again specifies a set of facts on a historical building (*El Cabildo*),

[1] http://www.w3.org/2009/08/skos-reference/skos.html

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 33–36
Paris, 10 November 2011

its location (*Buenos Aires*), and dedication date (*1933*), but the match between ontology label and text is not straightforward and requires the representation of linguistic information to compute morphological and syntactic variants.

2.2 Multilingual Knowledge Access

Ontology lexicalisation can be extended to multiple languages, enabling applications such as multilingual ontology-based question answering. Consider the following question in English, Dutch, German and Spanish:

- "Who painted the Mona Lisa?"
- "Wie schilderde de Mona Lisa?"
- "Wer malte die Mona Lisa?"
- "¿Quién pintó la Mona Lisa?"

Intuitively, the answer to these questions should be the same and thus independent of the specific language the question is expressed in. According to our main hypothesis, we claim that these questions could be translated into a normalized language-independent representation that can be evaluated with respect to semantically structured data. For example, we could use a formal query in the SPARQL language to express these questions in a way that abstracts from the original language:

```
PREFIX rdf: .../22-rdf-syntax-ns#
select ?who  where {
<http://dbpedia.org/.../Mona_Lisa>
<http://dbpedia.org/.../artist>
?who
}
```

The strings enclosed in angle brackets represent URIs (Uniform Resource Identifiers) that uniquely identify a certain entity (*Mona Lisa*) and a property (*artist*). The fact that the label of the property artist is English should not mislead; the URI represents a real-world relation between paintings and their creators and just happens to be labeled with an English string for the sake of human readability. The existence of such a relation is however independent of a specific language. In any case, in order to map the above question into a normalized and language-independent representation, i.e. the SPARQL query above, we require knowledge about the fact that the verb "schilderen" in Dutch, "malen" in German, "pintar" in Spanish and "paint" in English all refer to the property artist.

3 A Lexicon Model for Ontologies

Given the motivations for ontology lexicalisation given by the use cases outlined above and the fact that a solution for this seems missing in current state of the art research and best practices, we propose a formal model for the proper representation of the continuum between: i) ontology semantics; ii) terminology that is used to convey this in natural language; and iii) linguistic information on these terms and their constituent lexical units. As this model in essence enables the creation of a lexicon for a given ontology, we call this a *lexicon model for ontologies*.

3.1 Requirements

The requirements for a lexicon model for ontologies address several different goals. In particular, the model should: i) represent linguistic information relative to the semantics given by the ontology, thereby avoiding the representation of unnecessary lexical features that may lead to overgeneration of term variants; ii) strict separation of 'world knowledge' (describing domain objects that are referenced by lexical objects) from 'word knowledge' (describing lexical objects); iii) enable easy uptake of the model by providing a simple core model, supplemented with a set of modules that can be used, extended or ignored upon need.

3.2 lemon: lexicon model for ontologies

The proposed lexicon model for ontologies ('lemon') is described in detail in the 'lemon cookbook'[2]. Here we provide a summary of its most prominent features, starting with the lemon core, which is organized around a *core path* as follows:

- **Ontology Entity**: URI of an ontology element to which a **Lexical Sense** points, providing a possible linguistic realisation for that **Ontology Entity**
- **Lexical Sense**: functional object that links a **Lexical Entry** to an **Ontology Entity**, providing a sense-disambiguated interpretation of that **Lexical Entry**

[2] http://lexinfo.net/lemon-cookbook.pdf

- **Lexical Entry**: morphosyntactic normalisation of one or more **Lexical Form**
- **Lexical Form**: morphosyntactic variant of a **Lexical Entry**, including inflection, declination and syntactic variation
- **Representation**: standard written or phonetic representation for a **Lexical Form**

In addition, lemon has a number of modules that allow for further modeling:

- The **linguistic description module** is concerned with the use of data categories such as ISOcat for describing lemon elements. Although lemon itself is a meta-model and therefore agnostic as regards the specific data category set used, specific data categories can be used in particular instances of the lemon model.
- The **morphology module** is concerned with the analysis and representation of inflectional and agglutinative morphology. The module allows the specification of regular inflections of words by use of Perl-like regular expressions.
- The **phrase structure module** is concerned with the modeling of lexical entries that are syntactically complex, such as phrases and clauses, to enable representation of the syntactic structure of such lexical entries.
- The **syntax and mapping module** is concerned with a description of lexical 'predicates' (sub-categorisation frames with syntactic arguments) and semantic predicates (properties with subject/object) on the ontology side and the mapping between them.
- The **variation module** is concerned with a description of the relationships between elements of a lemon lexicon: sense relations (e.g. translation) require a semantic context, lexical variations (e.g. plural) require a morphosyntactic context, form variations (e.g. homographs) include all other variations.

4 Conclusions

In this paper we presented a motivation for ontology lexicalisation that builds on use cases, among others, in knowledge acquisition from text and multilingual knowledge access. We argued that the representation of a lexical level in ontologies, beyond the semantic and terminological level, is needed for a proper use of ontologies in applications and also serves in integrating the terminology level with the ontology level. No previously available model (e.g. (Gangemi et al.2003), (Farrar and Langendoen2003), (Reymonet et al.2007)) fulfills all the requirements for an ontology lexicalisation model. We therefore developed a model (lemon) for this purpose, of which we discussed some of its main features and directions in which it is currently used. Full details of the model and details of its use are described in other papers to which we refer the interested reader (Buitelaar et al.2009), (McCrae et al.2011), (McCrae et al.forthcoming).

Acknowledgements

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion2) and the CITEC excellence initiative funded by the EU and the DFG.

References

Buitelaar P. (2010) **Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions** *In: Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Oltramari, Alessandro Lenci, Laurent Prevot (eds.) Ontology and the Lexicon: A Natural Language Processing Perspective* Cambridge Studies in Natural Language Processing, Cambridge University Press.

Buitelaar P., P. Cimiano, P. Haase, M. Sintek (2009) **Towards Linguistically Grounded Ontologies** *Proceedings of the 6th European Semantic Web Conference.* Lecture Notes in Computer Science, Springer.

Farrar S., D. Terence Langendoen (2003) **A linguistic ontology for the Semantic Web** *GLOT International. 7 (3)*, pp.97-100.

Gangemi A,, R. Navigli, P. Velardi (2003) **The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet** *Proceedings of ODBASE*, Springer.

Guarino, N. (1998). **Formal Ontology in Information Systems** *In: N. Guarino (ed.) Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998.* IOS Press, pp.3-15.

McCrae J., D. Spohr, P. Cimiano (2011) **Linking Lexical Resources and Ontologies on the Semantic Web with Lemon** *Proceedings of the 8th European Semantic Web Conference*, Lecture Notes in Computer Science, Springer, Volume 6643, pp.245-259.

McCrae J., G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner (forthcoming) **Interchanging lexical resources on the Semantic Web** Accepted for publication in *Language Resources and Evaluation*, Springer.

Reymonet A., J. Thomas, N. Aussenac-Gilles (2007) **Modelling ontological and terminological resources in OWL-DL** *Proceedings of the ISWC07 workshop From Text to Knowledge: The Lexicon/Ontology Interface (OntoLex '07).*

Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction

Wiktoria Golik
MIG INRA UR1077
Domaine de Vilvert,
F-78350 Jouy-en-Josas,
France
wiktoria.golik
@jouy.inra.fr

Pierre Warnier
MIG INRA UR1077
Domaine de Vilvert,
F-78350 Jouy-en-Josas, France
LIG, Université Joseph Fourier, 385,
rue de la Bibliothèque, BP 53, F-
38041 Grenoble Cedex 9, France
pierre.warnier
@jouy.inra.fr

Claire Nédellec
MIG INRA UR1077
Domaine de Vilvert,
F-78350 Jouy-en-Josas,
France
claire.nedellec
@jouy.inra.fr

Abstract

The automatic population of a termino-ontology is a difficult and challenging task. We propose a text-based ontology extension method that was experimented and evaluated on, for a semantic annotation task in the biomedical domain. It is based on the linguistic analysis of terms and their heads. The head-based method improves both the identification of relevant areas of a termino-ontology and the matching of the corpus terms within these areas.

1 Introduction

Ontology population has seen much advancement in the past several years. It is an important area of research, in particular in the semantic annotation domain. There are several methods that specifically target ontology enrichment from text. All of these methods strive to be as automatic as possible. Some of them focus on the ontology structure in order to automatically infer semantic relationships (hyponyms, meronyms) (Euzenat, 2007). Others are supported by context analysis such as distributional semantics (Grefenstette, 1994) or patterns (Hearst, 1992).

Below we describe a text-based ontology extension method that we experimented on, for a semantic annotation task in the biomedical domain. Given a corpus and a domain specific termino-ontology, its primary aim is to identify the terms of the termino-ontology that are semantically close to the corpus terms. Further, it is used to annotate the corpus terms with the conceptual information provided by the termino-ontology. Conversely, the method will also be used for the automatic identification of relevant areas of the termino-ontology, which are semantically related to the given corpus terms. This mapping will be used for a semi-automatic extension of the selected areas in the termino-ontology with the corpus terms.

For relating corpus terms to termino-ontology terms, the method targets their internal structure. It therefore belongs to the class of linguistic methods based on the morpho-syntactic analysis of corpus terms (Jacquemin & Tzoukermann, 1999). The core of the method is based on the analysis of terms, their heads and the degree of head similarity (Hamon & Nazarenko, 2001). Our method is inspired by MetaMap (Aronson, 2001) which tags biomedical corpora with the UMLS Metathesaurus using syntactic analysis that takes into account lexical heads of terms.

The method has been successfully evaluated on the event extraction task of the BioNLP 2011 Bacteria biotope shared task (Bossy et al., 2011). The prediction of event arguments is done by automatically tagging corpus terms using termino-ontologies. Our linguistics-based approach achieves better results than shallow mapping methods.

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 37–39
Paris, 10 November 2011

2 Method

The BioNLP 2011 Bacteria biotope (BB) shared task consists of identifying bacteria and their locations in scientific documents. The locations belong to eight types to be predicted: Host, Host-part, Geographical, Food, Medical, Soil, Water and Environment. The participant system then has to relate bacteria to locations by a localization relation. Some locations (*e.g.* Geographical or Host) can be identified using named entity recognition. Contrarily, other types are more difficult to predict, since they can refer to any physical matter and are also subject to deep morpho-syntactic variations. They are noun phrases with adjectival and noun modifiers, verbal and prepositional complements. In order to overcome both the high degree of morphological variation and the incompleteness of the available lexicon, we experiment using a method based on the comparison of terms extracted from the corpus and termino-ontology terms. More precisely, the method identifies the semantically closest terms among them.

Our method applies to termino-ontology resources (TORs), defined as lexicalized ontologies. The terminological level consists of classes of canonical terms and synonyms with their syntactic properties. Each terminological class is attached to an ontology concept. We experimented using two different TORs of similar size, each containing approximately 1600 concepts: the Microorganism Biotope Termino-Ontology (MBTO) and the publically available EnvO habitat termino-ontology (Field et al., 2008). The first, focusing on bacteria biotope and phenotype modeling, has been previously developed at INRA. The second is more generalist, but it also targets habitat modeling. Given that the resources have not been created for the purpose of the experiment, we had to associate the BB task types to the MBTO and EnvO concepts. To do so, we took advantage of the hierarchical structure of the ontologies. We manually associated the high level nodes of the location hierarchies to the eight location types. The types of the lower level concepts were then automatically inferred. Local exceptions were manually handled. Importantly, the type assignment was consistent with the two TOR models: the concepts of the same type belong to contiguous areas.

The method relates corpus terms to TOR terms in two stages. The first stage extracts corpus terms using BioYatea. BioYatea is a version of YaTea (Hamon & Aubin, 2006), that we have extended and adapted to the biology domain. BioYatea provides information about the syntactic structure of terms, including the head and its modifiers. Head identification is a crucial point for our location identification and typing method. We consider that the head of a candidate location term is the most informative part and that it conveys the location type information. In most cases the term head is unambiguous with respect to the type.

At the second stage, the method assigns types to the candidate location terms produced by BioYatea. When both the corpus term and the TOR terms share the same head, the corpus term is assigned the type of the union of matching TOR terms. For instance, the corpus term *aquatic sediment* shares the head *sediment* with several TOR terms such as *lake sediment* or *spring sediment*. They all belong to the Environment area. Therefore, *aquatic sediment* acquires the type Environment. There are cases where the corpus term head appears in several type areas. For example, the head *spinach* from the corpus term *decayed spinach* belongs to three areas, namely Host (*diseased spinach*) Environment (*decayed spinach*) and Food (*cooked spinach*). A rule-based processing disambiguates among the several types by analyzing the term modifiers. Some heads may also be non informative and cannot be used to discriminate the type. These heads have been automatically detected among ambiguous TOR heads and recorded beforehand. If a candidate corpus term has such a head, we recursively search its subterms for an informative head. For instance, the head *environment* is non informative. Therefore, for the candidate term *cool soil environments*, the subterm *cool soil* is considered. Its head *soil* denotes the type Soil.

BioYatea extracted 1,873 candidate terms from the test corpus of the BB task. Table 1 details the number of candidate terms that the method associates to TOR terms with respect to the different typing strategies. There are two overall patterns. First, the well-adapted MBTO provides higher results for both strategies, exact match versus head match. Second, for both resources head-matching notably increases the term matching rate. This difference is especially marked for EnvO (91%). The results show that the head-matching strategy alleviates the TOR incompleteness, in particular for less adapted resources such as EnvO.

	MBTO	EnvO
% of corpus terms	16%	9%
Exact match	147	46
Main head of term	133	114
Subterm head	5	4
Ambiguous head	26	21
Total head matching	164 (52%)	139 (91%)
Total	311	185

Table 1: Number of terms typed using different-strategies.

The quality of the results is measured by the BB Task evaluation that provides entity recall measures per location type and the recall, precision and F-measure of the event prediction. The best event extraction F-measure (49%) is obtained using MBTO and the head-matching method. Using an exact match strategy with the same resource yielded a significantly lower F-measure (43.7%), with an entity recall of 64.3%. The head-matching improved the entity recall by 15 points, yielding 79.3%. This shows that the head based method is relevant for this task, suggesting that it is an appropriate and valuable strategy for relating semantically close terms from a corpus to an ontology.

3 Conclusion

The head-based method for the identification of semantically closer terms seems to be a good basis for a corpus-based termino-ontology extension. The use case from a biology domain presented above shows that using the term syntactic structure and head information can improve both the identification of relevant areas of a termino-ontology and the matching of the extracted terms within these areas. This method will be integrated in a user interactive application in order to assist ontology population, by suggesting probable areas for candidate term insertion. We also foresee the possibility of using such a method as the basis for a more automatic ontology population strategy. Beyond biology, we need to characterize the ontology and document genre the method is suitable for. We note that in the biology domain, physical locations are more likely denoted by terms with monosemous heads than with abstract and polysemous heads.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proceedings of AMIA Symposium 2001, 17-21.

Robert Bossy, Julien Jourde, Philippe Bessières, Maarten van de Guchte and Claire Nédellec. (2011). BioNLP Shared Task 2011: Bacteria Biotope. In Proceedings of Workshop Current Trends in Biomedical Natural Language Processing: Shared Task. Portland, USA.

Jerome Euzenat, Pavel Shvaiko. 2007. Ontology matching, Springer Verlag, Heidelberg, DE.

Dawn Field et al. 2008. Towards a richer description of our complete collection of genomes and metagenomes: the Minimum Information about a Genome Sequence (MIGS) specification. Nature Biotechnology, (26):541-547.

Gregory Grefenstette. 1994. Explorations in Automatic Thesaurus Discovery. Natural Language Processing and Machine Translation. Kluwer Academic Publishers, London.

Thierry Hamon and Sophie Aubin. 2006. Improving term extraction with terminological resources. In T. Salakoski et al. (Editors), Advances in Natural Language Processing 5th International Conference on NLP, (Fin- TAL'06):380-387. Springer.

Thierry Hamon and Adeline Nazarenko. 2001. Detection of synonymy links between terms: experiment and results. Recent Advances in Computational Terminology, 185-208. John Benjamins.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In A. Zampolli, (Editor), Proceedings of the 14 th COLING, 539-545, Nantes, France.

Christian Jacquemin and Evelyne Tzoukermann. 1999. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In T. Strzalkowski, (Editor). Natural language information retrieval. Text, speech and language technology, (7):25-74. Kluwer Academic Publishers.

Managing polysemy in the adventure tourism discourse with Frame Semantics

Isabel Durán-Muñoz
University of Málaga
iduran@uma.es

1 Introduction

This paper deals with polysemy in the adventure tourism domain, a sub-domain concerning the tourism sector with a high degree of polysemy. Our main goal is to manage this linguistic phenomenon by means of a methodology proposal based on semantic frames in order to reduce ambiguity in translation and terminological work.

Polysemy refers to the phenomenon that one word acquires different, though related, meanings, often with respect to particular contexts, that is, one term designating multiple concepts (Měchura, 2006). Polysemy poses quite difficult problems in terminography and other applied linguistics (like translation) regarding different aspects: representation of terminographical data in terminological resources (specialized dictionaries, glossaries, databases, etc.), structuring conceptual and terminological information, carrying out translations of polysemous units, among others.

In order to present this study, the main purposes of our paper are the following: firstly, to briefly discuss this new line of research in terminography and the assumptions of Frame Semantics; secondly, to introduce the main lexical features of the adventure tourism domain regarding polysemy and show several examples of polysemous units in this specialized domain according to the cases established; thirdly, to depict the steps to deal with this linguistic phenomenon in the domain under study, and, finally, to put forward some concluding remarks about the advantages obtained with this methodology.

2 Polysemy in the adventure tourism discourse

The adventure tourism terminology provides plenty of examples of polysemous units, which have been classified in the following three groups:

Case 1. A term which can be sorted under different conceptual categories (see Illustration 1). For example, “Hydrobob” is used both as an instrument and as an activity in the domain under study.

Case 2. A term which is linked to a conceptual category but which is used in different communicative situations and thus, presents specific features according to its related terms. For example, the term “Board” is employed as an instrument in a number of adventure activities but presents different features according to the specific activity it is used in (kitesurf, water skiing, windsurf).

Case 3. A term which refers to several concepts (and thus meanings) in one language but has different translation equivalents according to the several concepts it denotes. For example, the term “Kayak” in Spanish refers to two different concepts: <kayak> (instrument) and <hacer kayak> (activity) but in English the same instrument is called “Kayak” and the activity is “Kayaking.” This provokes a clear anisomorphism in the two languages at the terminological level.

As it is observed, polysemy is difficult to handle at a terminological level, since one unit refers to several concepts and meanings and, moreover, can have different translation equivalents. Likewise, it is difficult to represent this phenomenon on terminological resources (specialized dictionaries, databases, etc.).

3 Frame-based methodology

Our paper proposes a methodology to cope with polysemy from a conceptual level, that is, a methodology that takes advantage of the conceptual representation in order to facilitate the management of this linguistic phenomenon. Our methodology follows the most recent lines of ontology-based modern terminology research,

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 40–42
Paris, 10 November 2011

such as Termontography (Kerremans et al., 2003); Ontoterminology (Roche, 2009) or Ontoterminography (Durán-Muñoz, 2011/ forthcoming). Consequently, the proposed methodology is also corpus-based, descriptive, and systematic and in line with specialized lexicography.

These new lines of research consider traditional conceptual representations too limited to structure conceptual information as they just provide hierarchical categorizations based on generic-specific relations (IS_A) and part-whole relations (PART_OF). Likewise, they assume that knowledge representation needs a wider range of conceptual relations so as to provide greater coherence and specificity when structuring specialized domains. In this context, we uphold the inclusion of hierarchical relations, like traditional models, but also the need to include non-hierarchical relations, such as cause-effect and domain-specific relations (is_required_in). Ontologies (more specifically, domain ontologies) turn into a very valuable resource as they allow terminographers to build more complete categorizations and, thus, to carry out more suitable representations of specialized fields, as well as to handle polysemy and other linguistic phenomena.

Domain ontologies can be represented in different ways (linear, graphs and nodes, frames, etc.), each of which presents its own advantages and disadvantages. In our case, the frame-based methodology is applied as it is considered to be more suitable for clearly representing a communicative situation in which related concepts occur.

Frame-based terminology is a recent cognitive approach to terminography, which shares many of its assumptions with the Communicative Theory of Terminology (Cabré, 1999) and Sociocognitive Terminology (Temmerman, 2000) and is based on Fillmore's Frames (1976, 1982) and the cognitive models to represent knowledge and specialized domains.

The conceptual structure resulting from the application of frame semantics in terminology is similar to the representation of reality that humans create in our minds according to the neurolinguists (cf. Givón, 1995), since in both representations semantic relations are established between concepts that usually appear in the same communicative situation. For example, in the communicative situation of "buying a product", we usually match several concepts to this situation in our mind such as "seller", "buyer", "sell", "buy", "money", etc., which belong to the same semantic frame. [1] Therefore, it is asserted that in order to truly understand the meanings of units (both in general language and specialized discourse), it is required to first have knowledge of the semantic frames that underlie their usage, that is, the concepts and semantic relations established between them in concrete communicative situations.

Below an example of the application of frame-based terminology to create domain ontologies is provided.

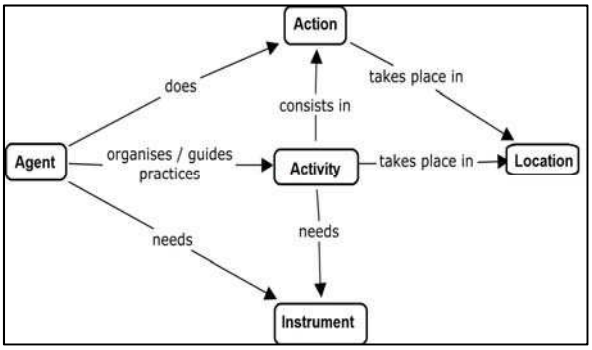

Illustration 1. Categorization of adventure tourism domain.

The Illustration 1 above is the result of a thorough analysis of the adventure tourism domain carried out by studying and managing the compiled specialized corpus and other information resources (dictionaries, legislation, etc.) supported by the assistance of domain experts. The categorization displays the prototypical situation of any event in the domain under study, that is, a conceptual template that provides the semantic frame according to its main categories and interrelations.

This prototypical situation comprises the five main categories detected in an initial phase: Agent, Activity, Action, Location and Instrument, which all are at the same conceptual level within the frame and all are considered necessary to understand the entire system. For example, to talk about an activity it is required to know who practices it, the place in which it takes place, the instrument needed and the action to be carried out. Consequently, the position that a concept occupies in a communicative situation is determined by the relations established with the other concepts included in this representation

[1] Petruck (1996: 1) defines frame as "any system of concepts related in such a way that to understand any one concept it is necessary to understand the entire system."

and, therefore, it eliminates any possible ambiguity at a language-independent level.

Once the frame-based categorization for the domain is been created, which could be modified (if necessary) or extended with further analysis, possible ambiguity is been reduced and almost eliminated. Subsequently, the next step is to manage the terminological level, where polysemy is encountered and needs to be handled.

In order to do so, the initial categorization is employed to classify terms and represent real communicative situations with the adventure tourism terminology. As a result, the terminological units are organized according to the frame-based representation and are easy to understand and differentiate from one to another and, also, to find translation equivalents in other languages.

4 Benefits of frame-based methodology

As stated above, adventure tourism terminology presents a high degree of polysemy, but thanks to the use of frame-based ontologies it is possible to deal with it by reducing the negative effects pertain to ambiguity, wrong translation equivalents, incomplete representation of domain, etc. The advantages of the application of Frame Semantics to deal with this phenomenon are manifold. Firstly, it provides a complete and coherent representation of the specialized domain categories and their interrelations within the same communicative situation at a conceptual level. Secondly, based on the conceptual level, it is easy to detect the different meanings attached to polysemous units taking their related concepts into account, that is, the meanings/concepts of a polysemous unit are distinguished thanks to the conceptual relations established with other concepts belonging to the same communicative situation. Consequently, it is possible to recognize the different meanings of a polysemous unit which can be placed under several conceptual categories (Case 1 above) or, also, clearly check the divergences of using the same unit in different communicative situations (Case 2 above). Thirdly, translation equivalents are easier to determine at a language-dependent level, as the language-independent level (conceptual level) is been properly structured and represents a prototypical communicative situation common to the working languages (Case 3 above). Therefore, each working language shares the same conceptual representation (or slightly adapted) and, as a consequence, it is easy to map them so as to find the suitable translation equivalents. And, finally, another remarkable advantage of employing semantic frames in terminology that is worthy to highlight is the possibility to systematically and coherently elaborate definitions based on the categories and the conceptual relations represented in the corresponding frame.

References

Cabré, 1999. La Terminología: Representación y Comunicación. Elementos para una teoría de base comunicativa y otros artículos. Barcelona: IULA. Universidad Pompeu Fabra.

Durán-Muñoz, I. 2011/forthcoming. La metodología del trabajo ontoterminográfico aplicado a la traducción. Berlin: Peter Lang.

Fillmore, C. 1976. Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280: 20–32.

Fillmore, C. 1982. Frame Semantics. In Linguistics in the Morning Calm, ed. The Linguistic Society of Korea, 111-137. Seoul: Hanshin.

Givón, T. 1995. Functionalism and grammar. Amsterdam/Philadelphia: John Benjamins.

Kerremans, K., Temmerman, R. y Tummers, J. 2003. Representing multilingual and culture-specific knowledge in a VAT regulatory ontology: support from the termontography approach. Lecture Notes in Computer Science, vol. 2889, 662-674.

Měchura, M. B. 2006. Finding the right structure for lexicographical data: experiences from a terminology project. Euralex 2006 - 12th Euralex International Congress, 6-9 September 2006, Turin, Italy.

Roche, C. 2007. Le terme et le concept: fondements d'une ontoterminologie. TOTh 2007 Terminologie & Ontologie: Théories et Applications. Annecy (France).

Temmerman, R. 2000. Towards New Ways of Terminology Description: The sociocognitive approach. Amsterdam/Philadelphia: John Benjamins.

A natural language ontology-driven query interface

Enrico Franconi and **Paolo Guagliardo** and **Sergio Tessaris**
KRDB Research Centre, Free University of Bozen-Bolzano, Italy
lastname@inf.unibz.it

Marco Trevisan
CELI Language & Information Technology, Torino, Italy
trevisan@celi.it

1 Motivations

Recent research showed that adopting formal ontologies as a means for accessing heterogeneous data sources has many benefits, in that not only does it provide a uniform and flexible approach to integrating and describing such sources, but it can also support the final user in querying them, thus improving the usability of the integrated system.

We introduce a framework that enables access to heterogeneous data sources by means of a conceptual schema and supports the users in the task of formulating a precise query over it. In describing a specific domain, the ontology defines a vocabulary which is often richer than the logical schema of the underlying data and usually closer to the user's own vocabulary. The ontology can thus be effectively exploited by the user in order to formulate a query that best captures their information need. The user is constantly guided and assisted in this task by an intuitive visual interface, whose intelligence is dynamically driven by reasoning over the ontology. The inferences drawn on the conceptual schema help the user in choosing what is more appropriate with respect to their information need, restricting the possible choices to only those parts of the ontology which are relevant and meaningful in a given context.

The most powerful and innovative feature of our framework lies in the fact that not only do not users need to be aware of the underlying organisation of the data, but they are also not required to have any specific knowledge of the vocabulary used in the ontology. In fact, such knowledge can be gradually acquired by using the tool itself, gaining confidence with both the vocabulary and the ontology. Users may also decide to just explore the ontology without actually querying the information system, with the aim of discovering general information about the modelled domain.

Another important aspect is that only queries that are logically consistent with the context and the constraints imposed by the ontology can be formulated, since contradictory or redundant pieces of information are not presented to the user at all. This makes user's choices clearer and simpler, by ruling out irrelevant information that might be distracting and even generate confusion. Furthermore, it also eliminates the often frustrating and time-consuming process of finding the right combination of parts that together constitute a meaningful query. For this reason, the user is free to explore the ontology without the worry of making a "wrong" choice at some point and can concentrate on expressing their information need.

Queries can be specified through a refinement process consisting in the iteration of few basic operations: the user first specifies an initial request starting with generic terms, then refines or deletes some of the previously added terms or introduces new ones, and iterates the process until the resulting query satisfies their information need. The available operations on the current query include addition, substitution and deletion of pieces of information, and all of them are supported by the reasoning services running over the ontology.

In this paper we summarise only the NL aspects of a tool based on those ideas, **Quelo**; for a complete picture of our ideas and of the tool refer to our papers (Franconi et al., 2011; Dongilli et al., 2004; Catarci et al., 2004; Catarci et al., 2005; Dongilli and Franconi, 2006; Franconi et al., 2010). Quelo relies on a web-based client-server architecture:

1. the tool logic, responsible of "reasoning" over the ontology in order to provide only relevant information w.r.t. the current query;

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 43–46
Paris, 10 November 2011

2. the natural language generation (NLG) engine, that given a query and a lexicalisation map for the ontology produces an English sentence; the lexicon is automatically generated from the ontology;

3. the user interface (GUI), that provides visual access to the query and editing facilities for it, allowing to interact with the reasoning sub-system while benefiting from the services of the NLG engine.

An online fully functional demonstrator of Quelo is freely accessible at:
`http://krdbapp.inf.unibz.it:8080/quelo/`

2 Natural language aspects

The natural language interface of the tool masks the composition of a precise query as the composition of English text describing the equivalent information need. Interfaces following this paradigm are known as "menu-based natural language interfaces to databases" or "conceptual authoring" (see, most notably, (Hallett et al., 2007)). As we have seen before, the users of such systems edit a query by composing fragments of generated natural language provided by the system through *contextual* menus. In (Franconi et al., 2010) we describe how the natural language rendering of a query is achieved.

We start by defining a particular linear form of the query that satisfies certain constraints, necessary to represent the elements of the query using a linear medium, that is, text. The constraints are enforced at the API level to ensure that different graphical user interfaces represent the query in a homologous way. Moreover, a consistent ordering of the query elements needs to be preserved during the operations for query manipulation to avoid confusing the end user. The linearised version of the query is then used as a guide for the language generation performed by the tool's NLG engine.

The natural language interface (NLI) of the tool relies on a natural language generation (NLG) system to produce the textual representation of the query, following an idea presented in (Tennant et al., 1983) and refined in (Hallett et al., 2007).

For the tool's NLI to work with a specific knowledge base (KB) a lexicon and a template map must be provided for it. To ease the burden of developing these resources from scratch, we let the system generate them automatically. The functionality we implemented allows to produce all the resources necessary to configure our NLI for use with a new KB, using as a source of data the ontology itself.

2.1 Natural Language Generation module

NLG systems use techniques from artificial intelligence and computational linguistics to produce human-readable texts out of machine-readable data. The Query Tool uses NLG to represent the whole query, along with all the elements that the user can use to refine it, as English text. The generated text is enriched with links that connect it to the underlying logical form of the query. This allows the user to operate on the query simply by editing an English text.

Unlike most NLG systems, ours is built to let the user determine the structure of the generated text by inserting, replacing and removing snippets of it. While in the classic NLG pipeline the information to be conveyed in the text and its order is determined by the document planning module, in the Query Tool it is the user who decides both the information to be displayed and its arrangement.

As the Query Tool is not tailored to any specific domain, its NLG module is simple enough to be adopted in any context and it is not bundled with all the resources that are needed to generate text out-of-the-box. Therefore, in order to use it on a specific knowledge base, the system must be provided with a *lexicon* and a *template map*. The former contains the words to be used in the generated text; the latter is the bridge between the natural language and the knowledge representation language, associating each concept/role name with a generation template. Each such template contains the syntactic and lexical information necessary to generate a fragment of text representing the associated concept or role.

We selected the syntactic features available in the templates, hence supported by the generator, in order to keep the system simple while still being expressive enough. For this purpose, we collected and analysed a corpus of more than 12.000 unique relation identifiers and we partitioned them according to the recurring syntactic patterns. For each class of the partition, we then proposed a common natural language representation template. The result of this study is a set of simple

but effective templates for representing most ontology relations using natural language.

During the first stage (*microplanning*) of the generation, linguistic information stored in the template map and in the lexicon blends with the logic information encoded in the query into a single structure, known in the NLG literature as *text specification* and consisting of a list of syntactic trees with inflected lexemes on its leaves. The NLG system operates on this structure to aggregate groups of adjacent syntactic structures into single more complex structures, and to select and replace existing referring expressions with more appropriate ones. These two tasks are known in the literature as *aggregation* and *referring expressions generation*, respectively. At the same time, the system keeps track of which element of the text specification is associated with which element (either a node tag or an edge tag) of the query. An association holds when the syntactic element is the result of the instantiation of a template associated with the element of the query. These associations are used for enriching the generated text with links to the underlying query.

The linearisation of the query simplifies the effort required by the referring expressions generation, as referring expressions that need to be reworked always appear in subject position. Our algorithm replaces a subject with a pronoun whenever the previous sentence had the same subject, otherwise the subject is left unchanged. Although ambiguous expressions may occur, ambiguity is not a crucial issue as these expressions originate from user operations upon a selected element, which always becomes the target of the referring expression. Our aggregation module performs simple aggregation tasks such as aggregating sentences with the same subject, eliding the subject and parts of the verb if it is feasible.

Once these operations are completed, the text specification is ready to be transformed into the final text. This task, known as *surface realisation*, produces a list of text tokens, some of which are connected to edge or node labels. This list is finally fed to the GUI, that displays it to the user.

Elements populating the menu for addition and substitution operations undergo a similar processing. To produce the textual representation of such an element, the system makes a temporary copy of the portion of query affected by the operation. The operation is carried out on this portion and the resulting structure is fed to the generation pipeline used for entire queries. The outcome of this process is the text which will appear on the menu.

2.2 Generation of lexicon and template map

For the tool's NLI to work with a specific knowledge base (KB) a lexicon and a template map must be provided for it. Devising these resources requires an understanding of both the domain of interest and basic linguistic notions such as verb tenses, noun genders and countability. We briefly describe here how the system can generate these resources automatically. This technique follows an approach to domain independent generation proposed in (Sun and Mellish, 2006), after the learning of a rich corpus of relations. The functionality we implemented allows to produce all the resources necessary to configure our NLI for use with a new KB, using as a source of data the ontology itself. It has to be noted that the process is not completely reliable, therefore system engineers must review the result and make the necessary corrections.

The idea is based on the observation that KBs already contain some form of linguistic information. In real-world ontologies, every concept and relation has a unique identifier (ID), which most of the times is not just an arbitrary string, but a mnemonic chosen by the knowledge engineer to describe the intended meaning of the identified concept or relation. Moreover, within these IDs, certain syntactic patterns occur more frequently than others.

In our approach, each relation ID is first tokenized according to an algorithm that takes advantage of the naming conventions used by ontology engineers. Second, the tokenized ID is fed to a custom part-of-speech tagger built around QTAG (Tufis and Mason, 1998). The resulting tagged tokenized ID is then lightly preprocessed before being finally passed to a transformation rule, chosen among thirteen different ones, that produces a template for the template map of the NLG system.

For the design of the transformation rules, we analysed our corpus, containing more than 12.000 relation IDs, in order to devise a partition of the domain in terms of syntactic patterns. The classes defined in this partition are s.t. to each relation of the same class can be applied a simple transfor-

mation in order to obtain a template. Each such transformation is also a uniform interpretation of the intended meaning of each relation ID in the class. Some care is needed when giving a uniform interpretation to syntactic patterns, as there are situations in which the same syntactic pattern is to be interpreted differently. For instance, the relation IDs "country_of_nationality" and "language_of_country" share the same syntactic structure, but the first relation should be read as "the country of nationality of X is Y", while the second as "the language of X is Y". Each of the thirteen rules we defined corresponds to one class of the partition, and together they can handle 93% of the relations of the average ontology.

The system has been formally evaluated with some ontologies (e.g., (Ordnance Survey, ; Drummond et al.,)), contributing 64 unique relations in total. It is now available online and it has been used in many different contexts. From the IDs of these relations we automatically generated relation templates, which were then inspected in order to evaluate their usability in text generation. The result of the evaluation revealed that for 42 out of 64 relations (65%) the generated template is suitable for direct use with the Query Tool's NLI. The result suggests that although the generation of the template map is not totally reliable, it is nevertheless useful in that it speeds up the work of systems engineers, as they do not need to create the whole map from scratch, but only have to review the generated map and repair eventual errors. This improves the portability of the Query Tool's NLI, making it faster and easier to switch to a different knowledge base.

References

Tiziana Catarci, Paolo Dongilli, Tania Di Mascio, Enrico Franconi, Giuseppe Santucci, and Sergio Tessaris. 2004. An ontology based visual tool for query formulation support. In *Proc. of the 16th Eur. Conf. on Artificial Intelligence (ECAI 2004)*.

Tiziana Catarci, Paolo Dongilli, Tania Di Mascio, Enrico Franconi, Giuseppe Santucci, and Sergio Tessaris. 2005. Usability evaluation tests in the SeWAsIE (SEmantic Webs and AgentS in Integrated Economies) project. In *Proceedings of the 11th International Conference on Human-Computer Interaction (HCII 2005)*.

Paolo Dongilli and Enrico Franconi. 2006. An Intelligent Query Interface with Natural Language Support. In *Proc. of the 19th Int. Florida Artificial Intelligence Research Society Conference (FLAIRS 2006)*, Melbourne Beach, Florida, USA, May.

Paolo Dongilli, Enrico Franconi, and Sergio Tessaris. 2004. Semantics driven support for query formulation. In *Proc. of the 2004 Description Logic Workshop (DL 2004)*.

Nick Drummond, Matthew Horridge, Robert Stevens, Chris Wroe, and Sandra Sampaio. Pizza ontology. The University of Manchester.

Enrico Franconi, Paolo Guagliardo, and Marco Trevisan. 2010. An intelligent query interface based on ontology navigation. In *Proc. of the Workshop on Visual Interfaces to the Social and Semantic Web (VISSW 2010)*, February.

Enrico Franconi, Paolo Guagliardo, Marco Trevisan, and Sergio Tessaris. 2011. Quelo: an ontology-driven query interface. In *Proceedings of the 24th International Workshop on Description Logics (DL 2011)*.

Paolo Guagliardo. 2009. Theoretical foundations of an ontology-based visual tool for query formulation support. Technical Report KRDB09-5, KRDB Research Centre, Free University of Bozen-Bolzano. `http://www.inf.unibz.it/krdb/pub/TR/KRDB09-05.pdf`, October.

Catalina Hallett, Donia Scott, and Richard Power. 2007. Composing questions through conceptual authoring. *Computational Linguistics*, 33(1):105–133.

Ordnance Survey. Great britain's national mapping agency. `http://www.ordnancesurvey.co.uk/oswebsite/ontology/`.

Xiantang Sun and Chris Mellish. 2006. Domain independent sentence generation from RDF representations for the Semantic Web. In *Proc. ECAI'06 Combined Workshop on Language-Enhanced Educational Technology and Development and Evaluation of Robust Spoken Dialogue Systems*.

Harry R. Tennant, Kenneth M. Ross, Richard M. Saenz, Craig W. Thompson, and James R. Miller. 1983. Menu-based natural language understanding. In *Proc. 21st Annual Meeting of the Association for Computational Linguistics*, pages 151–158. Association for Computational Linguistics.

Marco Trevisan. 2009. A portable menu-guided natural language interface to knowledge bases. Master's thesis, University of Groningen.

Dan Tufis and Oliver Mason. 1998. Tagging Romanian texts: a case study for QTAG, a language independent probabilistic tagger. *Proc. 1st Int. Conf. on Language Resources and Evaluation (LREC'98)*, pages 589–596.

Representing term variation in *lemon*

Elena Montiel-Ponsoda
Guadalupe Aguado-de-Cea
Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
{emontiel, lupe}@fi.upm.es

John McCrae
Semantic Computing Group
CITEC
Universität Bielefeld
jmccrae@cit-ec.uni-bielefeld.de

Abstract

In this contribution our objective is to define term variation, analyze the state of the art, and propose a new classification of term variants according to our representation purposes in *lemon*, a lexicon-ontology model to enrich ontologies with linguistic descriptions.

1 Introduction

A term variant has been defined as "an utterance which is semantically and conceptually related to an original term" (Daille et al., 1996). The same author expands this definition by explaining what is meant by *utterance*, *original term*, and s*emantically and conceptually related terms* (Daille, 2005). An *utterance* is an attested form encountered in a text. It is considered to be a variant with respect to an *authorised term*, i.e., a term listed in an authoritative terminological resource and accepted by a certain community. And it can be related to the original term in three forms: 1) by a synonymy relation, 2) by reflecting a "semantic distance from the reference term", or 3) by a conceptual link.

According to Daille (2005), the adopted definition of term variation depends on the purpose of the final application. For instance, in information retrieval the term variants usually handled are morpho-syntactic variants (*histamine of the wine* vs. *wine histamine*[1]) or variants related by a conceptual link (*printer* vs. *laser printer*).

In this contribution we concentrate on those variants that are considered synonyms and on those that reflect a "semantic distance" but that refer to the same concept. In doing so, we will not be dealing with those terminology variants related by means of a conceptual link. The reason for this is that we aim to analyze terminology variants with respect to an ontology or conceptual model, and we argue that conceptual relations will be already available in the knowledge model. However, we also foresee some mechanisms for the case that conceptually related variants are to be represented outside the ontology.

We understand **synonym related variants** as those term variants that are *semantically coincident but formally* different, as defined in Cabré (2008). With regard to **variants that reflect a semantic distance**, we include those variants that are *semantically and formally different* (Freixa, 2002; Cabré, 2008) but still refer to the same ontological concept.

In section 2 we propose a classification based on state of the art works and provide examples of each type of term variant. Then, in section 3 we describe how we aim at representing terminology variation in *lemon*, an ontology-lexicon model proposed in the framework of the Monnet project in order to linguistically enrich ontologies with lexical, terminological and syntactic information.

2 Typologies of variants revisited

Based on previous classifications of terminology variation (Freixa, 2002; Daille, 2005; Cabré, 2008) we identify two main groups of term variants: 1) term variants that are semantically coincident but formally different, and 2) term variants that are semantically and formally dif-

[1]Some examples have been extracted from Daille (2005) and Cabré (2008)

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 47–50
Paris, 10 November 2011

ferent. This has representation consequences as will be shown in section 3.

Group 1) would include,

- **graphical and orthographical variants** (*localization* vs. *localisation*);
- **inflectional variants** (*cat* vs. *cats*);
- **morpho-syntactic variants** (*nitrogen fixation* vs. *fixation of nitrogen*).

Regarding group 2), here we are dealing with terms that correspond to one and the same concept, but whose usage reflects a different aspect of the concept or a different intention on the side of the user, thus the semantic and formal distinction. This shows the pragmatic aspects necessary to be considered in scientific communication. It means that the use of one term or the other is conditioned by a certain cognitive intention and highlights certain dimensions or features of the concept that will make its use more appropriate in certain situations. This phenomenon has been termed *multidimensionality* (Broker, 1997; Rogers, 2004). As explained in Fernández-Silva et al., (2011), "multidimensionality occurs when a concept can be seen from more than one perspective and can therefore be classified and designated in more than one way based on the different characteristics that it possesses". In Cabré (2008) these term variants are also referred to as *partial synonyms*.

According to these definitions, we consider that the following term variants belong to this group:

- **stylistic or connotative variants** (*man* vs. *bloke*)
- **dialectal variants** (*gasoline* vs. *petrol*)
- **pragmatic or register variants** (*headache* vs. *cephalalgia*)
- **diachronic variants** (*tuberculosis* vs. *phthisis*)
- **domain or concept dimension variants** (*swine flu* vs. *pig flu* vs. *H1N1* vs. *Mexic pandemic flu*; *MRSA* (as Methicilin-resistance Staphylococcus aureus) vs. *HA-MRSA* vs. *CA-MRSA*)
- and what we dubb **explicative variants** (*immigration law* vs. *law for regulating and controlling immigration*).

It could also be argued that the term variants belonging to group 2) refer to different concepts, or, at least, to concepts belonging to different ontologies or to ontologies in the same domain created with different purposes. However, we claim that since they are pointing to the same concept or object in the world, they can be represented as term variants for that concept. In the context of our research we are able to capture these terminological variants in a complex model of lexical descriptions that is to be published with domain ontologies, namely, the *lemon* model (McCrae, 2011).

In *lemon*, concepts are represented by the ontology, and terms are associated with concepts by means of a principled link represented by the class `LexicalSense`. It is this intermediate class that allows us to capture those semantic properties of term variants that make them *semantically and formally distinct*. In the next section, we aim at illustrating the representation of term variants in *lemon*.

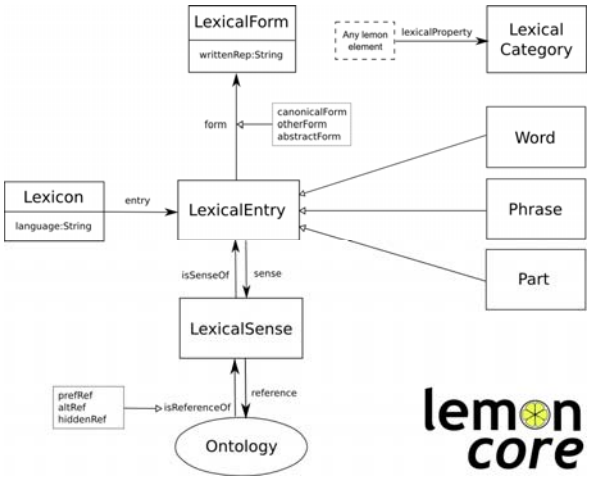

Figure 1. lemon core

1 Terminology variation in *lemon*

The core classes of the *lemon* model are the ones that make up the main path between the ontology and the lexical entry, its forms and written representations, as can be seen in Figure 1. Since concepts as defined in ontologies and lexical entries as defined in lexicons cannot be said to overlap, the *LexicalSense* class provides the adequate restrictions (usage, context, register, etc.) that make a certain lexical entry appropriate for naming a certain concept in the specific context of the ontology being lexicalized. This class will be key in making a distinction between those term variants included in group 1) and the ones included in group 2). Essentially, the main difference is that those terms variants considered *se-*

mantically coincident but formally different will be pointing to the same *LexicalSense*, whereas those considered *semantically and formally different* will be linked to different lexical senses, which in its turn are pointing to the same ontology element. Let us illustrate this with some examples.

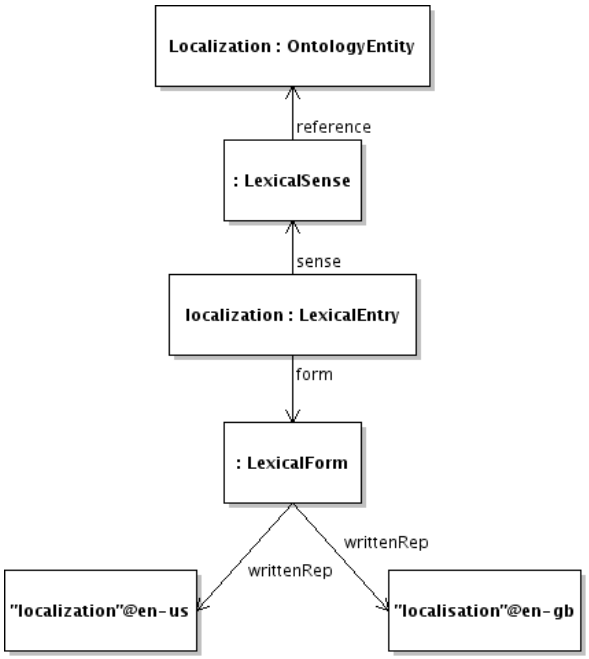

Figure 2. Example of orthographical variants

In Figure 2, we have included an example of the so-called graphical or orthographical variants. There we see that they are represented as two different written representations of the same *LexicalForm*, associated to the same *LexicalEntry* and pointing to the same *LexicalSense* and ontology concept. As these differences are only due to orthographical rules and not reflected in the spoken language, we consider them to be the same form of the entry.

In Figure 3 we represent two different lexical entries (*nitrogen fixation* and *fixation of nitrogen*) that are associated to the same *LexicalSense*, as their differences in format do not have any meaning or pragmatic consequences, but further represent the same meaning in the context of the ontology.

Figure 3. Example of morpho-syntactic variants

Finally, in Figure 4 we aim to illustrate one example of term variants which are semantically and formally different, in that they are used in different geographical settings. With the aim of capturing that restriction, we associate each *LexicalEntry* to a different *LexicalSense*, and account for that usage restriction.

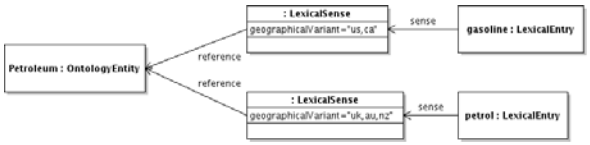

Figure 4. Example of dialectal variants

A similar approach would be valid for the rest of variants included in group 2).

Acknowledgments

This work is supported by the EU project Monnet (FP7-248458), the Spanish project BabeLData (TIN2010-17550), and the CITEC excellence initiative funded by the EU and the Deutsche Forschungsgemeinschaft.

References

Bowker, L. 1997. You say "flatbed colour scanner", I say "colour flatbed scanner": A descriptive study of the influence of multidimensionality on term formation and use with special reference to the subject field of optical scanning technology. Terminology 4(2):275-302.

Cabré, M. T. 2008. El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología (I). IBÉRICA 16:9-36.

Daille, B. 2005. Variations and application-oriented terminology engineering. Terminology 11(1):181-197.

Daille, B. Habert, B. Jacquemin, C and Royauté, J. 1996. Empirical observation of term variations and principles for their description. Terminology 3(2):197-257.

Fernández-Silva, S. Freixa, J. Cabré, M.T. 2001. A proposed method for analysing the dynamics of cognition through term variation. Terminology 17(1):49-73.

Freixa, J. 2002. La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient. PhD Thesis. Universitat Pompeu Fabra, Barcelona.

McCrae, J. Spohr, D., Cimiano, P. 2011. Linking lexical resources and ontologies on the semantic web with lemon. The Semantic Web: Research and Applications, 245-259.

Rogers, M. 2004. Multidimensionality in concepts systems: A bilingual textual perspective. Terminology 10(2):215-240.

Index

TIA 2011

9th International Conference on Terminology and Artificial Intelligence

Proceedings of the Workshops

TIA 2011 was supported by:

inalco
Institut national des langues et civilisations orientales

université PARIS DIDEROT

EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

afia
ASSOCIATION FRANCAISE POUR L'INTELLIGENCE ARTIFICIELLE

IAOA